

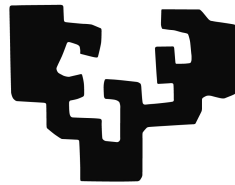
---

# Machine learning based anomaly detection for industry 4.0 systems

---

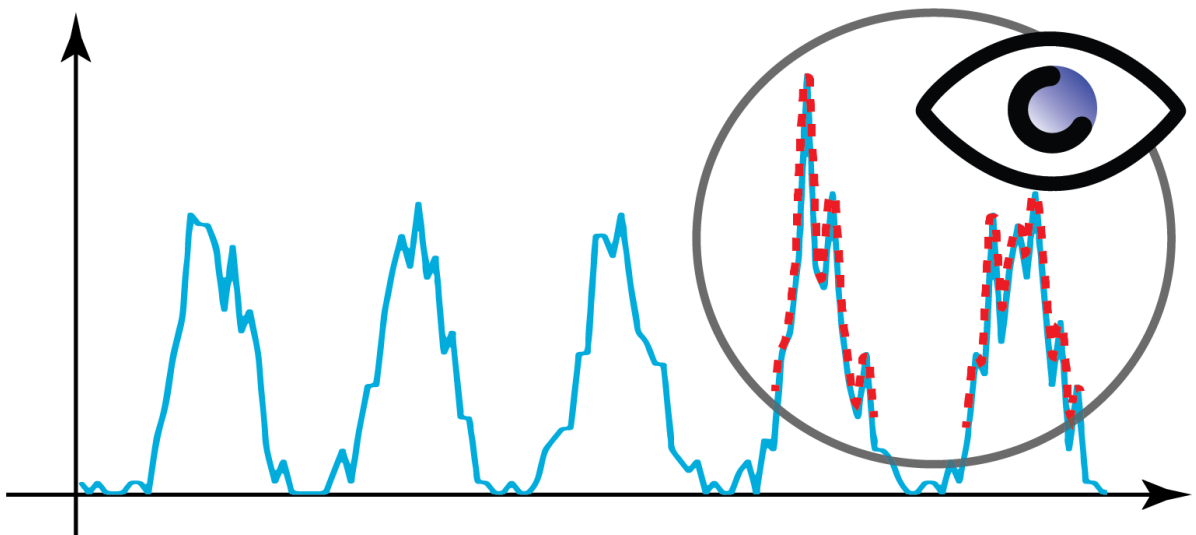
David Velásquez Rendón  
Universidad del País Vasco/Euskal Herriko Unibertsitatea  
January 2023

eman ta zabal zazu



Universidad  
del País Vasco

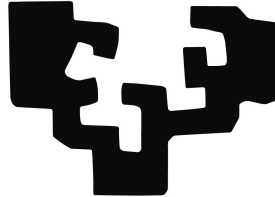
Euskal Herriko  
Unibertsitatea





UNIVERSITY OF THE BASQUE COUNTRY  
Faculty of Informatics

eman ta zabal zazu



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea

**Machine learning based anomaly detection for  
industry 4.0 systems**

Supervised by:  
Basilio Sierra Araujo  
and  
Mikel Maiza Galparsoro

Submitted by:  
David Velásquez Rendón  
For the academic degree of Doctoral Programme in Informatics  
Engineering

January 2023



---

## Acknowledgements

---

This thesis is dedicated to my father, who has supported me throughout all my studies, enabling me to achieve success in my life. I am deeply grateful for your reliable presence, your unwavering support, and for being a source of stability, joy, and strength in our family. You are my role model, and I will always carry you in my heart. My mother, who has also supported me with her unconditional love and encouragement, also deserves a special mention. Thank you for always being there for me. I would also like to extend my gratitude to my uncle Juan Carlos, who has played a significant role in my life, being like a second dad and always being there to lend a helping hand.

I am also thankful to EAFIT University, the University of the Basque Country, and the Vicomtech Foundation for providing me with the opportunity to undertake this thesis through their sponsorship and research environment. I express my appreciation to Ricardo Mejia for introducing me to Vicomtech and to Jorge Posada for giving me this opportunity. Additionally, I would like to thank Ricardo Taborda for his support and assistance from EAFIT University.

I am incredibly grateful to my tutors, Mikel Maiza and Professor Basilio Sierra, who provided me with the guidance, motivation and encouragement necessary to achieve my PhD goals. Their expertise in the field has been invaluable to me and I am so thankful for their sharing their knowledge with me. My appreciation also goes to Professor Mauricio Toro for his guidance from EAFIT University and for his contributions to the scientific aspects of my thesis, his support was fundamental in the completion of my research project. Their unwavering support, time and knowledge has made this project a success and I am deeply grateful for it.

I would like to extend my sincere appreciation to my friends, who have been there to support me during all the challenging moments of this process. Their encouragement and support have been invaluable in helping me to grow personally. In particular, I would like to thank Julian, Tony, Dider, Angelica, Laura, Camilo, and David.

I am also grateful to my colleagues at Vicomtech, especially those in the Data Intelligence for Energy and Industrial Processes department, who have provided me

with opportunities to share my experiences and enjoy coffee breaks together. I would like to extend a special mention to Juan Odriozola, who has not only been a friend but has also introduced me to the delights of Basque gastronomy. I would also like to thank Mikel Lopez for his help with work-related challenges and for his constant companionship.

Finally, I want to extend my gratitude to all my family for their support and for the wonderful times we have shared throughout this journey.

David Velásquez Rendón

January 2023

Donostia-San Sebastián

---

## Contents

---

<b>Acknowledgements</b>	<b>vi</b>
<b>List of Figures</b>	<b>xii</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Glossary</b>	<b>xv</b>
<b>Abstract</b>	<b>xix</b>
<b>I Body of the dissertation</b>	<b>1</b>
<b>1 Background</b>	<b>3</b>
1.1 General introduction . . . . .	3
1.2 Research environment . . . . .	4
1.2.1 University EAFIT . . . . .	5
1.2.2 Vicomtech . . . . .	5
1.2.3 Projects . . . . .	7
1.3 Structure of the dissertation . . . . .	9
<b>2 Motivation</b>	<b>11</b>
2.1 Boundaries . . . . .	11
2.2 Research questions . . . . .	13
2.3 Research outcomes . . . . .	14
<b>3 Theoretical concepts</b>	<b>17</b>
3.1 Machine Learning . . . . .	17
3.2 Ensemble Learning . . . . .	22
3.3 Anomaly Detection . . . . .	24

3.3.1	Statistical anomaly detection methods . . . . .	25
3.3.2	Classification anomaly detection methods . . . . .	25
3.3.3	Clustering anomaly detection methods . . . . .	26
3.3.4	Similarity-based anomaly detection methods . . . . .	26
3.3.5	Soft Computing anomaly detection methods . . . . .	27
3.3.6	Knowledge-based anomaly detection methods . . . . .	27
3.3.7	Combination Learners anomaly detection methods . . . . .	27
3.3.8	Time domain anomalies . . . . .	28
3.3.9	Frequency domain anomalies . . . . .	28
<b>4</b>	<b>Architecture</b>	<b>31</b>
4.1	Background . . . . .	31
4.2	Case study . . . . .	33
4.2.1	Smart-Water Case Study: Industrial Wastewater Treatment Plant “La Cartuja / EDAR 4.0 Project” . . . . .	34
4.2.2	Industrial Quality Testbench Case Study: Rotary Pneumatic Machines Company “MAPNER / EDAR 4.0” . . . . .	35
4.2.3	Smart IoT Embedded System Case Study: Rotary Pneumatic Machines Company “MAPNER / SISTELIA Project” . . . . .	37
4.3	Innovation . . . . .	38
4.4	Conclusions and future work . . . . .	40
<b>5</b>	<b>Data Acquisition</b>	<b>41</b>
5.1	Background . . . . .	42
5.2	Case study . . . . .	43
5.3	Innovation . . . . .	47
5.4	Conclusions and future work . . . . .	47
<b>6</b>	<b>Supervised Ensembling</b>	<b>49</b>
6.1	Background . . . . .	49
6.2	Case study . . . . .	52
6.3	Innovation . . . . .	56
6.4	Conclusions and future work . . . . .	58
<b>7</b>	<b>Semi-supervised Ensembling</b>	<b>59</b>
7.1	Background . . . . .	60
7.2	Case study . . . . .	61
7.2.1	Manufacturing-stage pipeline . . . . .	63
7.2.2	Operation-stage pipeline . . . . .	65
7.3	Innovation . . . . .	68
7.3.1	Manufacturing pipeline results . . . . .	68
7.3.2	Operation pipeline results . . . . .	71
7.4	Conclusions and future work . . . . .	71



<b>8</b>	<b>Frequency-based Anomaly Detection</b>	<b>73</b>
8.1	Background . . . . .	73
8.2	Case study . . . . .	74
8.3	Innovation . . . . .	77
8.4	Conclusions and future work . . . . .	79
<b>9</b>	<b>Visual Analytics</b>	<b>81</b>
9.1	Background . . . . .	82
9.2	Case study . . . . .	85
9.3	Innovation . . . . .	87
9.3.1	Water quality monitoring . . . . .	87
9.3.2	Water quality prediction . . . . .	90
9.3.3	WWTP Model Creation & Simulation . . . . .	90
9.3.4	WWTP Model Optimisation . . . . .	93
9.3.5	User’s validation . . . . .	94
9.4	Conclusions and future work . . . . .	96
<b>10</b>	<b>Conclusions and future work</b>	<b>97</b>
10.1	Conclusions . . . . .	97
10.2	Future work . . . . .	99
	<b>Bibliography</b>	<b>114</b>
<b>II</b>	<b>Appended Papers</b>	<b>115</b>
<b>11</b>	<b>Summary of the appended papers</b>	<b>117</b>
11.1	Paper 1 . . . . .	117
11.2	Paper 2 . . . . .	118
11.3	Paper 3 . . . . .	118
11.4	Paper 4 . . . . .	119
11.5	Paper 5 . . . . .	120
11.6	Paper 6 . . . . .	121
<b>12</b>	<b>Appended papers</b>	<b>123</b>



---

## List of Figures

---

2.1	4IR Pillars. . . . .	12
3.1	Supervised ML Algorithm Example. . . . .	18
3.2	Unsupervised ML Algorithm Example. . . . .	18
3.3	Semi-supervised ML Algorithm Example. . . . .	19
3.4	Autoencoder architecture. . . . .	20
3.5	Recursive Neural Network architecture. . . . .	21
3.6	CNN architecture. . . . .	21
3.7	DNN architecture. . . . .	22
3.8	Ensemble learning architectures. . . . .	23
3.9	SPC using $3\text{-}\sigma$ anomaly detection. . . . .	27
3.10	Anomaly Types. . . . .	28
3.11	Anomalies in a STFT graphical representation. . . . .	30
4.1	Reference Architectural Model for Industrie 4.0 (RAMI 4.0). . . . .	33
4.2	EDAR 4.0 architecture . . . . .	35
4.3	MAPNER Testbench architecture . . . . .	37
4.4	SISTELIA architecture . . . . .	38
4.5	4IR generic architecture . . . . .	40
5.1	Black box . . . . .	44
5.2	Final physical detailed design . . . . .	46
5.3	Final design of the system's cybernetic part . . . . .	46
6.1	CLR stages . . . . .	50
6.2	CLR ML pipeline . . . . .	54
6.3	Weighted average example . . . . .	56
6.4	Box plot results . . . . .	57
7.1	ML hybrid pipeline . . . . .	62
7.2	ML Manufacturing pipeline . . . . .	64

7.3	ML Manufacturing pipeline . . . . .	65
8.1	Frequency anomaly detection method . . . . .	75
8.2	Sample result spectrogram with the corresponding states. . . . .	76
8.3	Spectrogram Results . . . . .	78
9.1	Visual Analytics framework . . . . .	84
9.2	EDAR 4.0 architecture . . . . .	87
9.3	Visual Analytics Water Quality Monitoring Platform. . . . .	89
9.4	Visual Analytics Water Quality Prediction Platform. . . . .	90
9.5	Energy consumption model simulation . . . . .	91
9.6	Confusion matrix for electric model . . . . .	92
9.7	Variable influence for electric model . . . . .	92
9.8	Decision tree for electric model . . . . .	93
9.9	Energy consumption model optimisation . . . . .	94
9.10	Confusion matrix for water quality model . . . . .	95
9.11	Variable importance for water quality model . . . . .	96

---

## List of Tables

---

3.1	Techniques for anomaly detection . . . . .	24
5.1	Concept Scoring . . . . .	45
6.1	Hyperparameters and $F_1$ -score . . . . .	55
6.2	Submodel weights . . . . .	55
6.3	Final performances . . . . .	56
6.4	ANOVA table . . . . .	58
7.1	Air-Blowing Machines' data set characteristics. . . . .	66
7.2	Variables preprocessing at Manufacturing Stage. . . . .	67
7.3	Weights for the predictions of each submodel. . . . .	67
7.4	Machine A - Confusion Matrix (Test Set). . . . .	68
7.5	Machine B - Confusion Matrix (Test Set). . . . .	69
7.6	Machine C - Confusion Matrix (Test Set). . . . .	69
7.7	Machine A - Metrics table (Test Set). . . . .	70
7.8	Machine B - Metrics table (Test Set). . . . .	70
7.9	Machine C - Metrics table (Test Set). . . . .	70
7.10	Performance results of each model in microseconds. . . . .	71
8.1	Low pass digital filter inputs. . . . .	76
9.1	Water Quality requirements from European Directive 91/271/EEC . . . . .	82



- 4IR** 4th Industrial Revolution. xi, xix, 3, 4, 8, 17, 31–34, 36–40, 86
- AI** Artificial Intelligence. xix, 3, 4, 9, 31, 33, 38–40
- CBM** Condition-Based Maintenance. 59
- CLR** Coffee Leaf Rust. xi, 43, 44, 47–50, 52–58
- CNN** Convolutional Neural Network. 19, 21
- CPS** Cyber-Physical Systems. 3, 27, 31, 32, 45
- DL** Deep Learning. 19, 49, 52, 53, 56–58
- DNN** Deep Neural Network. xi, 19, 22, 25
- DSP** Digital Signal Processing. 28
- FFT** Fast Fourier Transform. 29
- IoT** Internet of Things. 3, 4, 8, 27, 31, 32, 34, 37–40, 42, 44, 45, 86
- k-NN** k Nearest Neighbours. 26
- LOF** Local Outlier Factor. 26, 60, 64, 67, 71
- LSTM** Long Short Term Memory. 20
- ML** Machine Learning. xi, xii, 7–9, 17–19, 22, 25, 34, 49, 51, 54, 59–65, 67, 71, 72, 83
- NN** Neural Networks. 27

**OCSVM** One-Class Support Vector Machine. 60, 64, 67, 71

**PDS** Product Design Specification. 43, 44

**RNN** Recursive Neural Network. 19, 20

**RS** Remote Sensing. 42, 43, 45, 47, 49–53, 57, 58

**SCADA** Supervisory Control And Data Acquisition. 3, 11, 34, 86, 87

**VA** Visual Analytics. 4, 8, 81

**WSN** Wireless Sensor Networks. 42, 43, 45, 47, 49, 51–53, 57, 58

**WWTP** Waste Water Treatment Plant. 7, 8, 34, 81, 82, 84, 86–88, 91, 93, 94, 96







The Fourth Industrial Revolution has brought new disruptive technologies that are gradually being implemented in industries. These technologies range from the Internet of Things, Artificial Intelligence, 3D Printing, and Augmented Reality, among others. These enable improving existing industrial processes, “digitising” many operations that were previously carried out manually, and adding intelligence to the information. However, the 4IR presents significant challenges in integrating these technologies into existing legacy systems or creating new innovative products and services that possess these new technologies. One step in the transition to 4IR is using sensors on machines to capture data on what is happening around the system context. This data can then be sent to a computer system in the cloud, where AI-based technologies can be used to analyse and gain insights into system events, e.g. the occurrence of system anomalies. Detecting anomalies can proactively help the early detection of possible failures in processes and machines, and foresee that they may occur at a specific time in the future, which aids improving production and maintenance processes.

This dissertation focuses on the study of anomaly detection in industrial systems, aiming to provide tools tested and validated in real use cases. This thesis aims to verify the applicability and implementation of 4IR technology architectures, AI-driven machine learning systems and advanced visualisation tools for supporting the decision-making processes based on the detection of anomalies. The topics investigated in this work range from the initial conception of a 4IR system based on a generic architecture, the design of a data acquisition system for subsequent analysis and modelling, the creation of ensemble supervised and semi-supervised models for anomaly detection, the detection of anomalies through frequency analysis, and the visualisation of the associated data using Visual Analytics.

The results obtained from each tool and approach developed in this thesis indicate that the proposed methodology for integrating anomaly detection systems in new or existing industries is valid. It is proved that the integration of 4IR architectures, ensemble machine learning models and Visual Analytics tools significantly enhances the anomaly detection processes for industrial systems. Furthermore, this work presents a guiding framework for data engineers and end users.

# Part I

## Body of the dissertation



This chapter will first present a general introduction describing Industry 4.0 and its derived technologies that are part of this work. It will then describe the research environments where the work was carried out, and finally the structure of the dissertation will be presented.

### 1.1 General introduction

The 4th Industrial Revolution (4IR) is an integration of systems and technology, where virtual and physical manufacturing systems work together flexibly and globally. Nevertheless, it is not just networked, intelligent systems. The scope is broader, from gene sequencing to nanotechnology, from renewable energy to quantum computing. These technologies' convergence and interplay across physical, digital, and biological domains distinguish the Fourth Industrial Revolution from its predecessors [130]. The 4IR poses new challenges to traditional industrial processes. This means either improving existing processes or creating new ones that use new technologies efficiently and exploit their full potential. In an increasingly competitive market, 4IR can be viewed as a disruptive innovation that positively impacts various industrial sectors by integrating new enabling technologies. Examples of these technologies include 3D printing, the Internet of Things (IoT), Cyber-Physical Systems (CPS), Artificial Intelligence (AI), Big Data, Robotics, Nanotechnology, and Quantum Computing [155, 78].

Improvements in internet speed, coverage, and bandwidth allow 4IR systems to process large amounts of data using cloud computing [102]. These large amounts of data come from many different sources, such as sensors, surveillance and data acquisition systems (SCADA), or third-party data sources, such as weather stations. Once data is collected, AI algorithms process these large amounts of data to provide additional knowledge to optimise processes and increase profitability. However, some

applications require real-time processing by using embedded system processing or edge computing technology to speed up the response to the 4IR system [133].

Regarding the data collection from processes, the IoT is nowadays used. IoT refers to the use of intelligently connected devices and systems leveraging data acquired by embedded sensors and actuators in machines and other physical objects [50]. Furthermore, these data are intended to be collected and stored in a cloud, which can then be analysed by different AI algorithms, as mentioned above.

Machine Learning algorithms provide a way to analyse these previously collected data. Their basis is a set of inputs (features), the model to be trained, and outputs (targets) [95]. A model is trained to predict outputs based on the inputs (features). Machine Learning methodologies can be differentiated into three main categories: (i) Supervised Learning, (ii) Unsupervised Learning, and (iii) Semi-supervised Learning. If the ground truth of the outputs is known, supervised algorithms are preferably used [125]. In cases where the output is not known, unsupervised methodologies can be used [54]. There are some cases where some outputs are known a priori or can be inferred, and for these, semi-supervised algorithms can be implemented [158].

One of the challenges that often exists in the industry is the detection of anomalies in their systems and processes. The definition of an anomaly is a moment in time in which the system's behaviour differs significantly from its previous normal behaviour [22, 2]. For example, a fluctuation in the turbine rotation frequency of a jet engine may indicate an imminent failure due to an anomaly. There is also the possibility of an anomaly indicating a positive trend; for example, many web clicks on a new product page may imply a high demand for the product. Notably, anomalies in data can provide insight into abnormal behaviour that can be translated into potentially helpful information in both cases. Anomaly detection can be used in several application areas, which can include intrusion detection, fraud prediction, failure detection in industrial equipment, and disease detection [15].

The above algorithms and methods usually need to be brought to an end user, so some visualisation or interaction is necessary. For the latter, Visual Analytics (VA) can be used. VA is a combination of automated analytic techniques with interactive visualisations to understand, reason, and make sound decisions based on large and complex amounts of data. VA focuses on creating new tools that allow users to: i) aggregate information to gain new insights from massive heterogeneous datasets, ii) detect the current state of the system and explore possible new states, iii) provide real-time feedback and take actions based on these responses [69].

This work will guide through all parts of the process, from data acquisition, through data analysis, and finally, visualisation, to present Machine Learning algorithm strategies for anomaly detection in Industry 4.0 systems.

## 1.2 Research environment

This section describes the research environments and some of the projects in which this work has been done.

### 1.2.1 University EAFIT

This research was done in collaboration with the University EAFIT (Colombia), in the Information and Communications Technologies Research, Development and Innovation Group (GIDITIC), which belongs to the Department of Informatics & Systems Engineering. GIDITIC implements research and development projects; offers advisory and consultancy services; and relies on partnerships, technical cooperation, and active participation in national and international research networks. The GIDITIC research group has the following research lines:

- **Agrotech.**
- Scientific computing and high performance computing.
- Ubiquitous computing.
- Digital content.
- Information and knowledge.
- Educational informatics, networks and virtual communities.
- Educational informatics innovation models.
- Educational informatics collaborative work.
- Educational informatics intelligent tutorials.
- ICT infrastructure.
- Software engineering and formal methods.
- Mixed reality and video games.
- Information security.
- Autonomous systems for decision making.

The agrotech line investigates projects focused on developing technologies focused on agriculture. For example, this line has developed projects focused on using machine learning to predict crop pests for remote crop monitoring.

### 1.2.2 Vicomtech

The Vicomtech Visual Interaction and Communication Technologies Centre Foundation is an applied R&D&I centre set up in 2000 and located in the Donostia-San Sebastian Technology Park (Spain). It currently has 19 top-level business and institutional partners in different fields related to its activity.

Vicomtech is part of BRTA (Basque Research and Technology Alliance), established by the Basque Government, SPRI, Regional Governments and Technological



Centres, which has the main function of responding to the technological and industrial challenges in the Basque Country and improving awareness of the centre at international level. It is also part of the international research network Graphics-Media.net (now GraphicsVision.AI), integrated by prestigious international applied research centres totally aligned with Computer Graphics and Multimedia technologies, which also gives it an active and strategic profile of internationalisation of its research activity.

Vicomtech is one of the few centres that simultaneously holds the internationally recognised UNE 166002:2014 and ISO 9001:2008 certificates, which place it at the forefront of quality in research and demonstrate its commitment to the quality of its processes. It has also recently obtained the Silver A from Euskalit in recognition of its Advanced Management, as well as the Quality Innovation of the Year 2015 Award (1st prize) for the best European Social and Healthcare Innovation. Recently, Vicomtech has achieved the recognition of the European Commission HR Excellence in Research that accredits its commitment to open, transparent and merit-based recruitment of researchers (OTM-R: Open, Transparent and Merit-based Recruitment of Researchers).

Since its foundation in 2000, its main function has been to carry out applied research in the area of interactive computer graphics and multimedia technology, focusing its areas of interest on:

- Industry and Advanced Manufacturing.
- Digital Media.
- **Data Intelligence for Energy & Industrial Processes.**
- Intelligent Systems for Mobility and Logistics.
- Digital Security.
- Intelligent Security Video Analytics.
- Speech and Natural Language Technologies.
- Digital Health and Biomedical Technologies.
- Connected & Cooperative Automotive Systems.

This project was carried out in the Data Intelligence for Energy & Industrial Processes department. In this department, Data Intelligence technologies are researched. They focus on collection, distribution, storage and especially the analysis of data in complex contexts, intending to discover characteristics, trends, relationships and, ultimately, the data's non-evident knowledge. This is especially relevant in Big Data environments, and any context in which the data obtained can improve the understanding of critical processes in a particular domain. Some modern Artificial Intelligence techniques, such as Machine Learning or Visual Analytics, are relevant

in specific Data Intelligence applications. In the discrete manufacturing and continuous processes industries, there is a very high unexploited potential for underlying knowledge, which can be obtained by applying Data Intelligence to already existing data. Furthermore, in data intelligence, historical scenarios and temporal series of high complexity are gathered (in some variables, in sample frequency, in a volume of information, in a type of data, in heterogeneous sources), as well as data produced in real-time, to show the domain experts patterns and trends allowing them to anticipate faults (e.g., preventative and prescriptive maintenance) or to improve their production strategies.

### 1.2.3 Projects

The following are the research projects the present PhD candidate has been working on during the doctoral thesis. These projects allowed for the experimentation of this thesis.

#### **Coffee Leaf Rust (Gobernación de Colombia, Ministerio de Ciencias 2017-2021)**

Coffee Leaf Rust has been a fungal epidemic disease affecting Colombian coffee trees since the 1980s. At the national level, it has caused massive defoliation, and in extreme cases, it has resulted in devastating losses of 70% to 80% of the harvest. In this regard, early detection of critical patterns that may indicate the presence of the disease would contribute strategically to reducing crop damage and increasing farmer profitability. Several studies have demonstrated the advantages of using technological methods to identify those patterns. As a result, in this study, a technological integration of Remote Sensing, Wireless Sensor Networks, and Deep Learning was created to evaluate its performance in diagnosing the Coffee Leaf Rust development stage in the Colombian *Caturra* variety. To that end, this project developed a cyber-physical data collection system that automatically collects data from a scale coffee crop and transfers it to a remote server via the Internet, as well as a ML-based diagnostic model. The results analysis revealed that the disease diagnosis made by visual inspection versus the proposed technological integration are statistically comparable.

#### **EDAR 4.0 (Hazitek Estratégico 2016-2020) & Digital Twin EDAR 4.0 (DFG 2021-2022)**

Estación Depuradora de Aguas Residuales 4.0 (EDAR 4.0 by its acronym in Spanish) is a research project aiming to create tools for optimising the operation and energy management of wastewater treatment plants (WWTPs). Water and energy engineering companies, process automation companies, WWTP equipment manufacturing companies, research centres, and universities all collaborate on different aspects of the project, with the ultimate goal of developing a cloud-based web platform integrating a complete set of tools to support the intelligent operation of WWTPs. The project's foundation is the collection of plant-wide data on all of the processes that

comprise a WWTP. These processes can be divided into three main standard sub-processes: i) the inflow process, which mainly represents the input of influent water and its pretreatment and primary treatment; Usually done in primary septic tanks or sedimentation ponds. ii) Biological treatment processes are the core part of the so-called secondary treatment and constitute the plant's primary waste water purification process. It is driven by different types of bacteria and protozoa and can be complemented by additional chemical treatments. iii) the drainage process; This is primarily the discharge of wastewater directly into the receiving waters or through a secondary septic tank or sedimentation tank, which is considered part of the post-treatment of the facility. With the data mentioned above, an IoT infrastructure that can be reached through the internet was developed, thus the overall WWTP, ICT infrastructure has to have (a secure) access to it. Several services, such as multiple plants management, cloud-based IoT data acquisition and storage, information monitoring (visualisation), data analysis and associated services such as Visual Analytics (VA), plant simulation, and plant optimisation through machine learning (ML), are integrated into such a cloud-based ICT infrastructure.

### **SISTELIA (Hazitek Competitivo 2019-2020) & RICVAS (Hazitek Estratégico 2021-2023)**

SISTELIA project acronym translates to “Intelligent Services for Industrial Blowers based on Digitalisation Technologies and Artificial Intelligence”. RICVAS translates to “Remote Integrated Control, Visualisation and Analytic Services”. The primary purpose of projects SISTELIA and RICVAS was to create and implement a cloud-based data management platform based on a 4.0 embedded system named “MAPNER Panel Control” (MPC), designed for the real-time acquisition, analysis, and visualisation of data and enriched information of machines around the world. The MPC includes an ad hoc Human Machine Interface (HMI) that collects data from an integrated, real-time data acquisition (DAQ) system that is integrated with sensors and provides real-time information to maintenance operators both directly on the machine (local visualisation) and via a 4G network Communications Interface. SISTELIA's architecture comprises a physical blowing machine, a 4IR embedded system, and a cloud data management platform. In these projects, real-time anomaly detection models were developed using Machine Learning techniques to improve preventive maintenance and enable predictive maintenance for the customer's MAPNER air blower machines.

### **RAPID (Gobierno Vasco 2020-2021)**

RAPID is a project launched by the Basque Government through the Basque Government's Department of Economic Development and Infrastructures (SPRI Group) to boost the Basque industrial sector by maintaining productive activity in the face of the COVID-19 pandemic danger (PRAP Euskadi).

The RAPID mobile phone application (“Rapid-App” app) within the project collects voluntary contact tracing information between employees of an organisation

during the workday to take rapid action in the event an employee shows symptoms or has tested positive for COVID-19.

In this project, Visual Analytics tools were developed in conjunction with proximity data analysis to assist the organisation in managing the pandemic.

## 1.3 Structure of the dissertation

This dissertation is divided in two parts: Part I contains the body of the dissertation; Part II includes a summary of this thesis' most relevant papers that are later included.

Part I is organised in nine main chapters, including this introductory chapter. A brief description of the content of each chapter follows next.

**Chapter 1:** Background. This chapter starts with a general introduction. Then, the research environment is described introducing the research centres and the main projects where this research work has been based on. Finally, the structure of the thesis is detailed.

**Chapter 2:** Motivation. The main context of the research is presented, together with the research questions and outcomes.

**Chapter 3:** Theoretical concepts. This chapter presents definition and properties of the theoretical concepts which are used in the dissertation.

**Chapter 4:** Architecture. In this chapter, a generic architecture for AI-driven Industry 4.0 systems is presented.

**Chapter 5:** Data Acquisition. A cyber-physical data collection system design for Coffee Leaf Rust is detailed.

**Chapter 6:** Supervised Ensembling. This chapter presents a method to detect anomalies through ML ensembling of different sensor sources with a case study of detecting the severity of Coffee Leaf Rust.

**Chapter 7:** Semi-Supervised Ensembling. A semi-supervised ensemble anomaly detection method based on Industry 4.0 is presented in this chapter.

**Chapter 8:** Frequency Anomaly Detection. This chapter presents a method to detect fractures in compacted hygroscopic materials by means of frequency analysis for anomaly detection.

**Chapter 9:** Visual Analytics. A Wastewater Treatment Plant Visual Analytics tool is explained in this chapter.

**Chapter 10:** Conclusions and future work. Finally, this parts ends with general conclusions and future work of the dissertation.



This chapter presents the boundaries within which the research problem of the current project is addressed. Then the research questions to be answered are presented in detail. Finally, the scientific contributions made in this project are listed.

## 2.1 Boundaries

The fourth industrial revolution brings new technologies that enable industries to improve their productivity and optimise their processes. These technologies include Big Data and Analytics, Autonomous Robots, Advanced Simulation, Cyber-Physical Systems Integration, Industrial Internet of Things, Cybersecurity, Cloud Computing, Additive Manufacturing and Augmented Reality. Figure 2.1 presents these pillars of the 4IR.

However, the transition of these legacy industrial processes and the creation of new industrial processes with 4IR-enabled systems pose a challenge for industries. Some of these challenges are listed below:

- Implementation of cloud services for industrial data: Industrial systems usually use programmable logic controllers (PLCs) for process monitoring and control. These are usually interconnected via industrial networks to a central system called SCADA (Supervisory Control And Data Acquisition). From this central, SCADA system, operators can monitor and control all processes. However, these systems are often limited in computing capabilities to the physical infrastructure. This presents the challenge of implementing services in the cloud, where more scalability and computational power can be made available [47].
- Ubiquitous process monitoring and control using emerging web technologies: In this case, industries require greater ease of access to their processes from

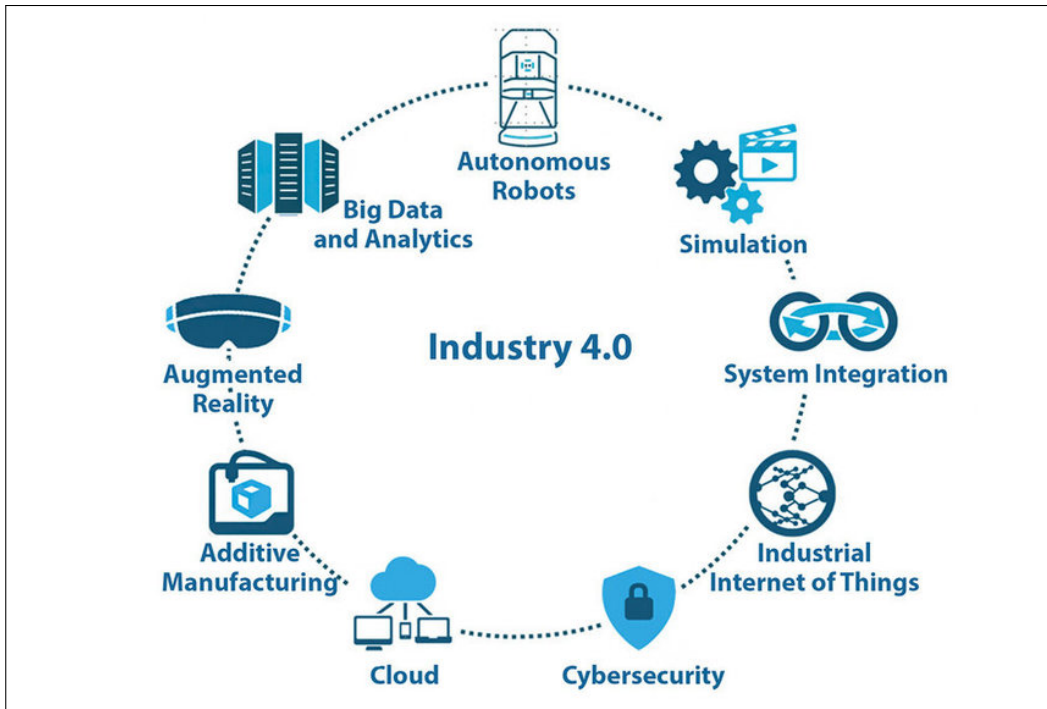


Figure 2.1: 4IR Pillars. Different emerging technologies available from the 4Th Industrial Revolution [109]

anywhere and securely. It is also necessary to make it easier for operators to understand the processes of the entire industry [26].

- Optimisation of computing capabilities for the appropriate selection of technologies such as Edge Computing and Cloud Computing: Industrial processes generally generate large amounts of data, so it is required to evaluate which processes can be processed by Edge Computing or Cloud Computing [133].
- Failure prediction in processes/products/machinery: Industrial processes have used fault detection techniques that sometimes do not account for the whole context of the process. This makes it necessary to analyse data from one or more processes to understand the system better [44].
- Integration of Big Data technologies in real-time for industrial data: Critical industrial processes require a real-time or near real-time response to avoid process failure. Therefore, the challenge is incorporating Big Data technologies to process this data in real-time and deliver an appropriate corrective action [92].
- Cybersecurity: Connecting industrial processes to the cloud can carry significant risks in terms of the security of the information collected [81].

In terms of fault detection, it is necessary, first of all, to have appropriate system data. Therefore, methods for acquiring data from multiple process sources, such as sensors and -existing process records, must be explored. Then, a model must be

created capable of predicting and identifying possible anomalies in the system. This model must consider all previously acquired variables and correctly decide which ones to use. From this comes the challenge of defining which method(s) are optimal and how they can be combined to obtain a more reliable answer. Finally, these results must be presented to the end user in a clear, concise and understandable way so they can make appropriate decisions in case of a possible failure in their industrial system. This presents challenges ranging from data acquisition, through the creation and combination of intelligent models, and finally, to deliver this information to the user in an easily interpretable manner. The research challenge of this project is to contribute to the prediction of failures integrated into remote data acquisition processes and the monitoring process through Visual Analytics.

## 2.2 Research questions

This thesis focuses on failure prediction for industrial systems, for which research centre Vicomtech has related real industrial projects. The problem relies on how to approach a process of anomaly detection in the field of the 4IR. For this reason, and based on the context of this research work, the following research question is posed:

**To what extent is it possible to implement machine-learning models and visualisation systems that allow the determination of anomalies on new Industry 4.0 systems?**

In order to obtain comprehensive answers, it is considered convenient to analyse different industrial use cases, which provide a scalable, applicable, robust solution. Based on the main research question, the following derived research questions are proposed:

1. What is the recommended design architecture for creating or adapting an industrial process with 4IR-enabling technologies?
2. How should industrial data be collected to enable the creation of intelligent models that can predict failures in 4IR systems?
3. How should data from different sensor sources be integrated into the models generated for failure prediction?
4. How should Machine Learning models be assembled to generate a more reliable prediction?
5. How should the results of the predictive analysis be delivered to the users so that they can understand them easily and make decisions based on them?

This work will attempt to answer these research questions through different use cases in different industrial contexts.



## 2.3 Research outcomes

The results of this research present an answer to the main research question: anomaly detection by using ensembling methods based on model performance. Additionally, processes related to the definition of 4IR architectures and the integration of these Machine Learning and visualisation technologies in the process management of these industries are addressed. The following articles, including scientific contributions from the authors, have been published in academic papers generated as part of this thesis to answer the previously mentioned research questions.

Journal articles:

- (a) Velásquez, D., Sánchez, A., Sarmiento, S., Toro, M., Maiza, M., & Sierra, B. (2020). **A method for detecting coffee leaf rust through wireless sensor networks, remote sensing, and deep learning: Case study of the Caturra variety in Colombia.** Applied Sciences (Switzerland).
- (b) Velásquez, D., Perez, S., Mejía-Gutiérrez, R., & Velásquez-López, A. (2020). **Crack Detection Method in Transport of Hygroscopic Particulate Compressed Material.** International Journal of Mechanical & Mechatronics Engineering, 20, 26–33.
- (c) Velásquez, D., Sánchez, A., Sarmiento, S., Velásquez, C., Toro, M., Montoya, E., Trefftz, H., Maiza, M., & Sierra, B. (2021). **A Cyber-Physical Data Collection System Integrating Remote Sensing and Wireless Sensor Networks for Coffee Leaf Rust Diagnosis.** Sensors.
- (d) Velásquez, D., P'erez, E., Oregui, X., Artetxe, A., Manteca, J., Mansilla, J. E., Toro, M., Maiza, M., & Sierra, B. (2022). **A Hybrid Machine-Learning Ensemble for Anomaly Detection in Real-Time Industry 4.0 Systems.** IEEE Access, 10, 72024–72036.
- (e) Velásquez, D., Vallejo, P., Toro, M., Odriozola, J., Moreno, A., Naveran, G., Maiza, M., & Sierra, B. (2023). **EDAR 4.0: Visual-Analytics for Waste Water Management.** IEEE Transactions on Industrial Informatics, 1-10. (Submitted)
- (f) G. Olaizola, I., Bruse, J. L., Odriozola, J., Artetxe, A., Velásquez, D., Quartulli, M., & Posada, J. (2023). Visual Analytics platform for Centralized Covid-19 Digital Contact Tracing. IEEE Computer Graphics and Applications, 1-14. (Accepted)

Conference articles:

- (a) Cestero, J., Velásquez, D., Suescún, E., Maiza, M., & Quartulli, M. (2022). Pysurveillance: A Novel Tool for Supporting Researchers in the Systematic Literature Review Process. In K. Nakamatsu, R. Kountchev, S. Patnaik, J. M. Abe, & A. Tyugashev (Eds.), Advanced Intelligent Technologies for Industry (pp. 239–248). Springer Nature Singapore.

- 
- (b) Velásquez, D., Toro, M., Bruse, J. L., Oregui, X., Maiza, M., & Sierra, B. (2022). **A Novel Architecture Definition for AI-driven Industry 4.0 Applications**. Proceedings - 2022 11Th International Conference on Industrial Technology and Management (ICITM), 1–7. (Accepted)

According to the above, only the most significant publications about the author (remarked in bold) and related to research questions will be included in Part II.



This chapter presents the theoretical concepts related to Machine Learning for Anomaly Detection in 4IR systems.

### **3.1 Machine Learning**

One of the emerging technologies proposed by the 4IR is Machine Learning (ML). ML is a field that studies the creation of methods that can learn from data, similar to how a human brain works in the learning process [35, 121]. These trained systems can then be used automatically to make decisions, such as identifying image elements, filtering emails (spam or non-spam), predicting the weather, and detecting anomalies in industrial systems, among others [134]. ML requires a set of data to be trained to learn the patterns or behaviour of the data and then perform a specific function. In the case of industrial systems, these have sensors that capture data on the physical variables of the process, which can then be stored in a database for the subsequent creation of ML algorithms [3].

Machine Learning algorithms can be categorised according to the availability of ground truth in the training data: supervised, unsupervised, and semi-supervised [120].

Supervised algorithms are those where a label is available for the training data. For example, there are images of coffee leaves with healthy and diseased labels [62]. Then an ML algorithm is trained to learn to recognise diseased leaves, as can be seen in Figure 3.1.

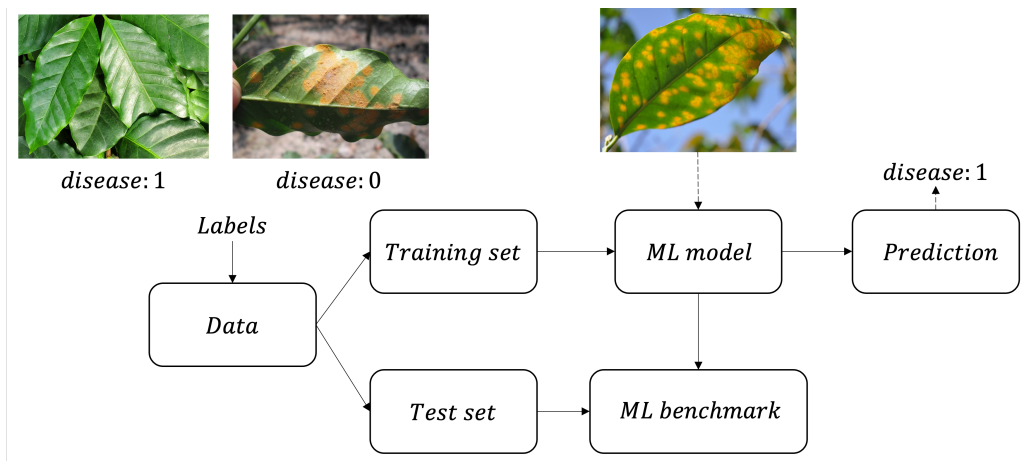


Figure 3.1: Supervised ML Algorithm Example. A pipeline of a supervised ML algorithm with a coffee disease classifier example.

In the case of unsupervised ML algorithms, there is no label as such but techniques can be used to group the data, for example, with clustering algorithms [33]. One of the most common techniques for unsupervised ML is the k-nearest neighbours algorithm, where the input data is grouped with a label based on its nearest neighbours. The most common metric for calculating distance between points is the euclidean distance, which is shown in Equation 3.1.  $d_{(x,y)}$  is the distance between a point  $x$  and  $y$ ,  $k$  is the number of classes.

$$d_{(x,y)} = \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (3.1)$$

Figure 3.2 displays an example of a clustering ML algorithm which groups data based on three different clusters.

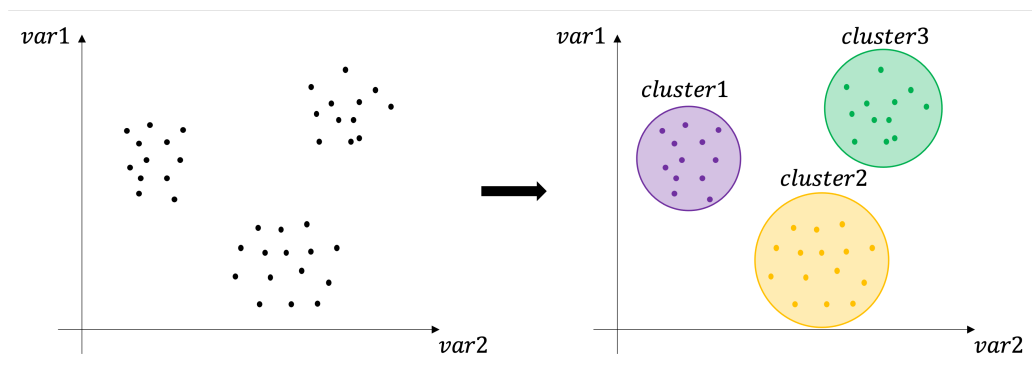


Figure 3.2: Unsupervised ML Algorithm Example. A clustering ML algorithm.

Semi-supervised ML algorithms are those with a small amount of labelled data and many unlabelled data remaining in the training set [158]. These are usually the case in industrial systems, where there is considerable sensor data, but the process

or machine's condition (label or ground truth) is unknown. Figure 3.3 displays the pipeline process for a semi-supervised ML algorithm. This pipeline first consists of a dataset with no labels or a small number of labels. Then there is a first Machine Learning model that can also be an algorithm to correctly classify the existing data labels and thus generate a pseudo-labelling training set. Finally, this labelled dataset is used to train the Machine Learning model and make the relevant predictions.

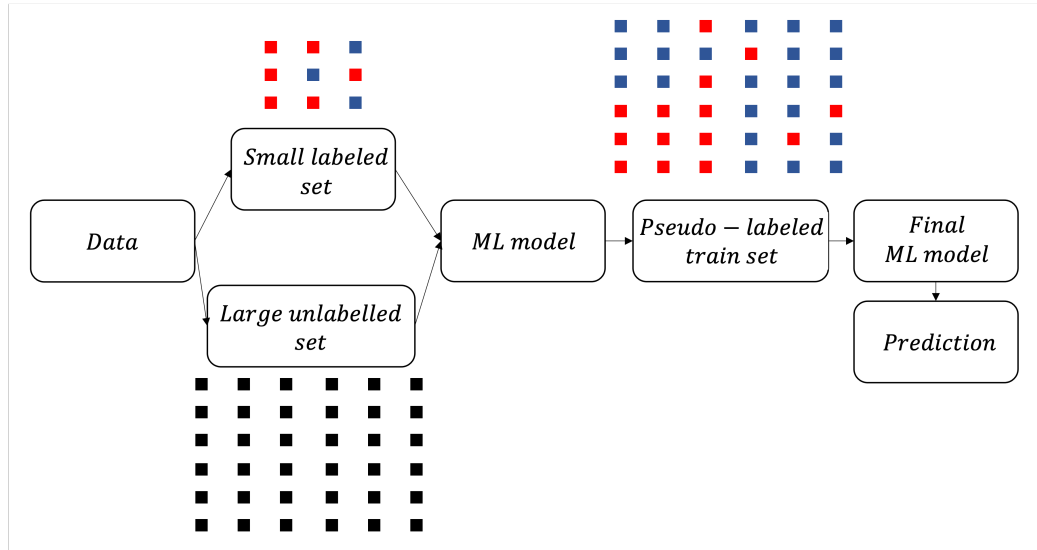


Figure 3.3: A pipeline of a semi-supervised ML algorithm.

Deep Learning (DL) is a branch of ML that allows automatic learning tasks to be performed more accurately and with greater identification capabilities than a traditional ML algorithm. These have been implemented and developed recently due to the rapid growth of Cloud Computing. The most common DL algorithms are Autoencoders, Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Deep Neural Networks (DNN) [96]. Autoencoders are unsupervised algorithms that learn to reconstruct the input layer data at its output layer. To achieve this, they first “encode” the data in their hidden layer, thus reducing its dimensionality, and then “decode” it until the original dimensions of the input layer are recovered in the output layer (see Figure 3.4).

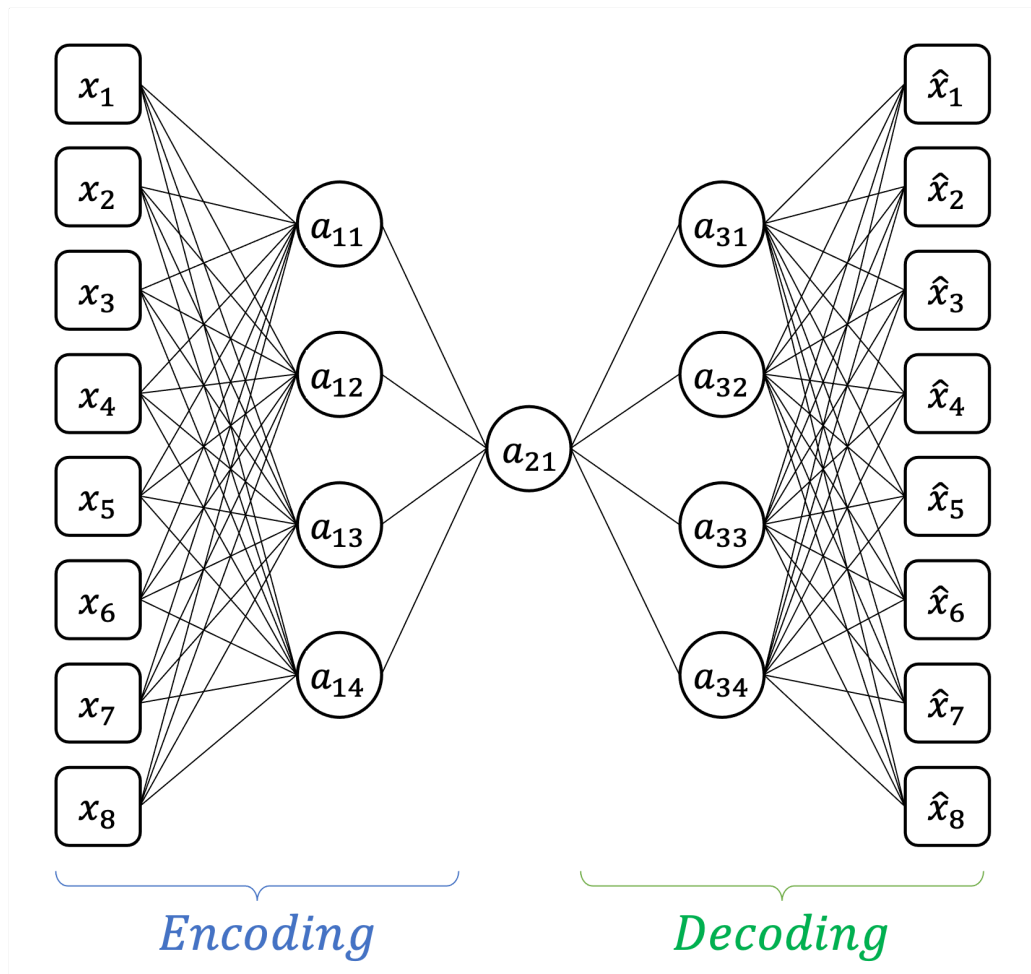


Figure 3.4: An autoencoder architecture.

RNNs work well for sequential data. They have an input layer, hidden layers and an output layer. RNNs differ from a traditional neural network in that they “remember” previous outputs, which are ultimately used as input for computation between neurons in their inner layers (see Figure 3.5). The most widely used RNN algorithm is Long Short Term Memory (LSTM), conceived by Hochreiter and Schmidhuber [55] in 1997, which is commonly used for speech recognition.

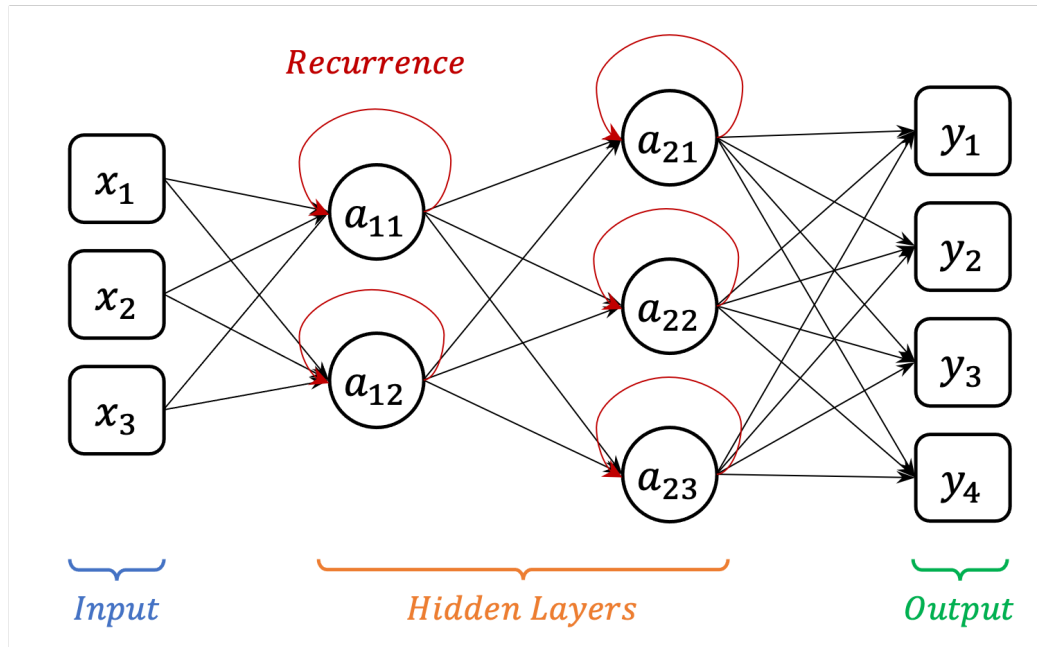


Figure 3.5: Recursive Neural Network architecture.

Convolutional neural networks (CNNs) have an input layer, a convolution layer, a pooling layer, a fully connected layer and an output as shown in Figure 3.6. The convolution layers allow abstracting the input data from the image into a feature map (also called an activation map). The pooling layers are responsible for reducing the size of the data by combining the outputs of the neuron clusters into a single layer with a single neuron in the next layer. Fully connected layers connect each neuron in one layer to all neurons in the next layer. CNNs allow the complexity of the entire network to be reduced by sharing the weights and reducing the number of neurons required through a pooling operation [74]. CNNs are commonly used for image processing but can also be used for time series [86], natural language processing [82], and recommender systems [71].

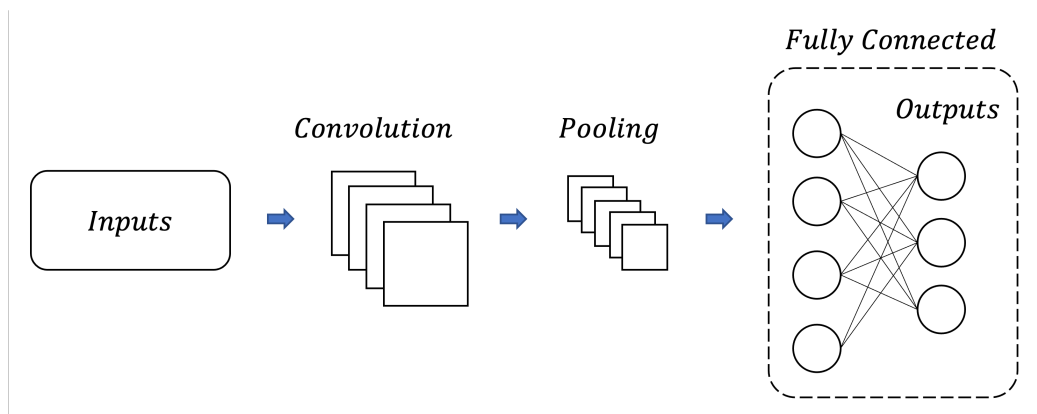


Figure 3.6: CNN architecture.



Deep Neural Networks (DNN) have an input, output, and multiple hidden layers (see Figure 3.7). They can learn patterns from the input data to predict an output. Such networks are in essence a fully-connected neural network. DNNs are commonly used in image or speech recognition [58].

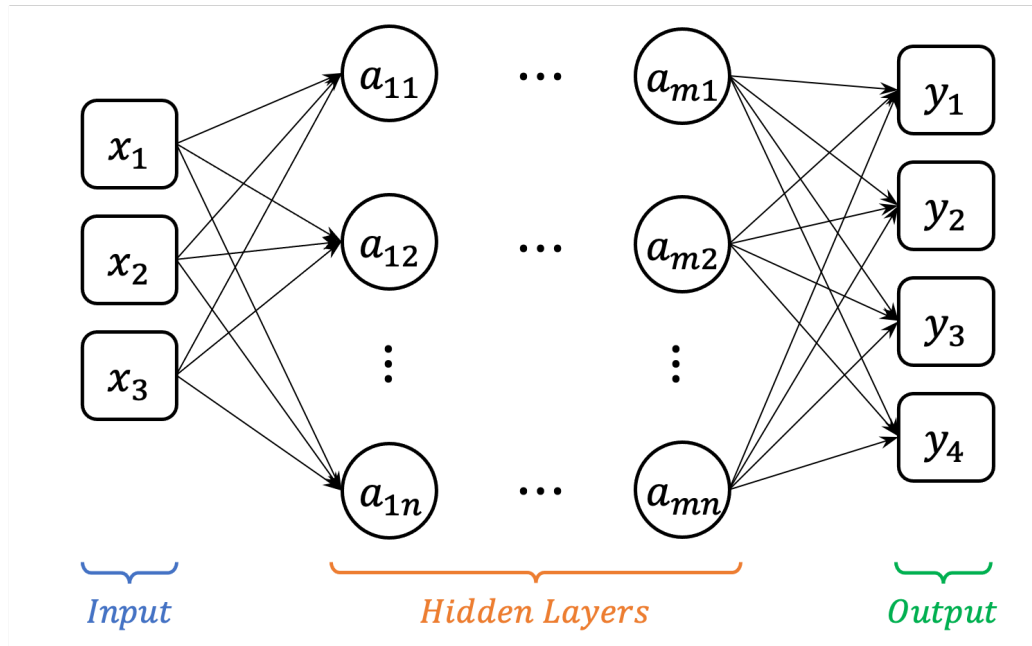


Figure 3.7: DNN architecture.

## 3.2 Ensemble Learning

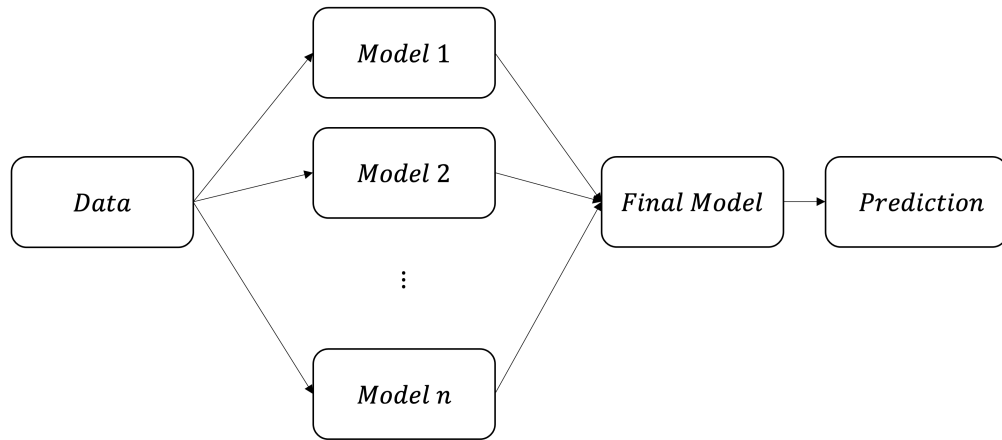
Ensemble learning is an ML methodology that seeks to improve the performance of models by combining several models. There are three main methods for ensembling: i) stacking, ii) boosting, and iii) bagging [126].

The *stacking* method consists of fitting different types of ML models (heterogeneous ML algorithms) using the same dataset and using another ML model to learn the best combinations of the models. Among the benefits of ensembling by stacking is the ability to leverage the capabilities of well-performing models to a greater extent than each model separately.

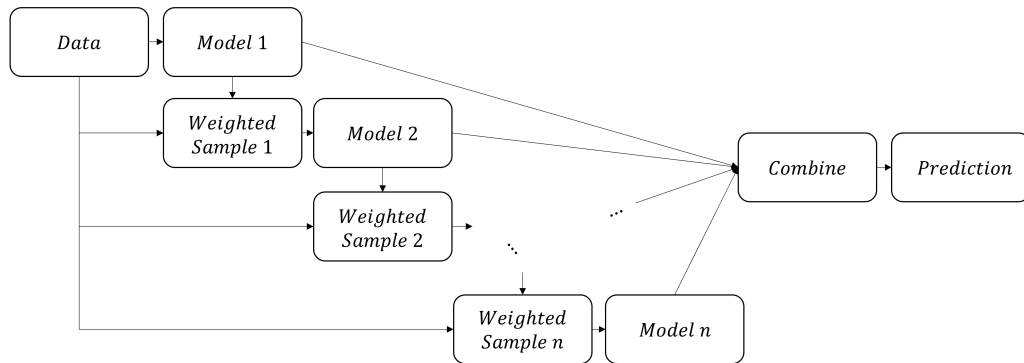
In *boosting*, models are added to the output of an initial model sequentially to correct or improve the predictions made by the previous models.

*Bagging* consists of training several decision trees (homogeneous ML algorithm) on different samples of the same dataset and then averaging their predictions.

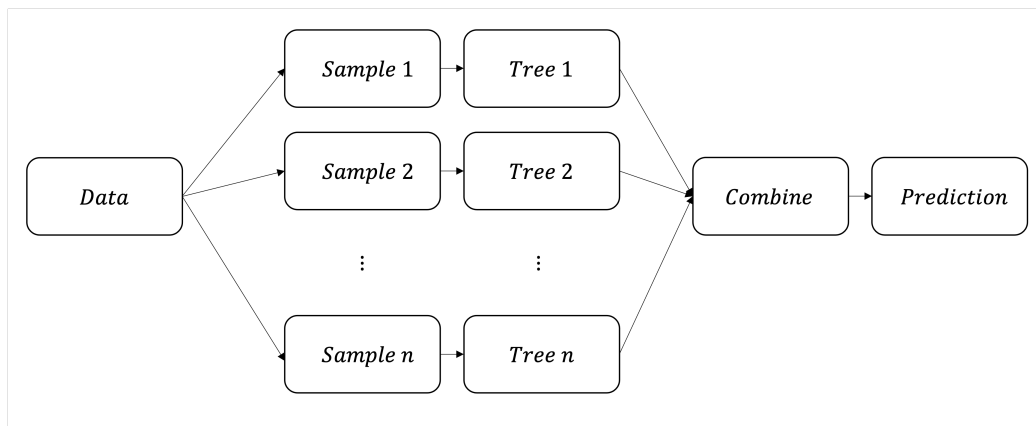
Figure 3.8 shows the architectures for each ensembling learning models.



(a) Stacking



(b) Boosting



(c) Bagging

Figure 3.8: Three different ensemble learning architectures.

### 3.3 Anomaly Detection

According to Hawkins [53], an anomaly can be defined as an observation that deviates significantly from other observations to such an extent as to raise suspicion that it was generated by a mechanism other than the system itself. Barnett et al. [12] define it as an observation (or subset of observations) inconsistent with the rest of the data. Anomaly detection methods can be divided into supervised, semi-supervised and unsupervised. Which method to use usually depends on the presence or absence of a label describing the anomaly. Labels can be categorical, for example, a binary label such as “abnormal behaviour” (1) or “normal behaviour” (0). Or, they can be numeric, for example, a value of “abnormal score” that goes from 0 (“normal”) to 1 (“completely abnormal”). Anomaly detection can be posed as a supervised learning task, but this is generally not the case, as there is often little or no data flagged as anomalous behaviour [21].

There are different techniques for anomaly detection, including statistical, classification, clustering, similarity-based, soft computing, knowledge-based and combination learners as summarised in Table 3.1. Furthermore, anomalies can be identified in the time domain or, in some cases, may be easier to see in the frequency domain. All of these categories are detailed below.

Table 3.1: Summary of techniques for anomaly detection[119]

<b>Technique</b>	<b>Sub Techniques</b>	<b>Examples</b>
Statistical	Parametric, Non-parametric	Box-plot, Grubbs test, Chi-square, PCA, Kernel methods
Classification	One Class, Multi-Class	Neural Networks, Bayesian Networks, SVM, Decision Trees
Clustering	Parametric, Non-parametric	DBSCAN, Rock, SNN, K-means, EM, LOF variants
Similarity based	Continuous and categorical data	k-NN variants, Relative Density
Soft Computing	GA, NN, Fuzzy and Rough Sets, Ant Colony	GANIDS, NN, DNN, CNN
Knowledge based	Rules and Expert Sys- tems, Ontology and Logic-based techniques	Decision Trees
Combination Learners	Ensemble based, Fusion based, Hybrid	Bagging and Boosting

### 3.3.1 Statistical anomaly detection methods

Statistical techniques adjust a predefined distribution to a given data and apply statistical inference to determine whether an instance belongs to that model. Instances with low probability are registered as anomalies [56]. The two typologies used by this method are parametric and non-parametric. The first makes assumptions about the underlying distribution of the data. Although the second method is somewhat less efficient at detecting anomalies, it is preferable because, a priori, it does not determine the structure of the model inferred from the data. The most common parametric methods are divided into gaussian and regression models. If a non-parametric approach is to be followed, such a classification can be based on histograms or kernels. Statistical methods are suitable for simple structured data with small sizes and quantities. A variety of methods can be used in such cases [119], including box-plots, Bayesian networks, autoregressive methods (autoregressive integrated moving average - ARIMA, autoregressive moving average - ARMA), Principal Component Analysis (PCA), ML-based methods, the Blum Floyd Pratt Rivest Tarjan (BFPR) algorithm, medcouple test, Grubbs test, and, comparison of distributions (QQ chart, Kolmogorov-Smirnov test, Kruskal-Wallis test, Wilcoxon signed range test).

### 3.3.2 Classification anomaly detection methods

Classification-based anomaly detection methods perform two main steps called training and testing. During the training phase, the system learns from available samples and creates classifiers. The model's performance is measured in the testing phase by trying patterns the classifier does not recognize. Depending on the labels available for training, the classifier can be divided into two categories: i) single class and ii) multiple classes. Example methods of single- and multi-class classifiers are neural networks, Bayesian networks, support vector machines (SVMs), fuzzy logic, and decision trees. These methods also work well in high-noise environments as discussed in [135, 38, 1, 157]. The advantage of classification-based methods is that they can distinguish between observations belonging to different anomalies (rather than a general class called "anomaly"), and test instances are compared to a predefined model so that their testing phases are fast [20]. However, classification methods rely on the ability to assign labels to different normal and abnormal classes, which is a difficult task. In addition, these methods assign labels to the test data, which can be a disadvantage when anomaly scores are desired. Classification-based methods can also classify by type of anomaly. Radial Base Functions (RBFs), SVMs, and derivatives are commonly used for single anomalies. RBF is very accurate and fast, especially for the controlled classification of individual anomalies. DNNs, induction rules, and decision trees are used for multiple anomalies. DNNs can provide exceptional recognition speed in static scenarios but can create problems for data that changes over time.

### 3.3.3 Clustering anomaly detection methods

Clustering methods are usually divided into two stages. First, the data are grouped using a clustering algorithm, and then the degree of variance is analysed according to the results obtained by the clustering [15]. There are some preliminary considerations for data instances of these unsupervised methods. On the one hand, regular data samples belong to the global cluster. On the other hand, anomalies do not belong to any particular cluster. In addition, normal data samples are located near the nearest cluster centroids, while anomalous data are further away. Finally, normal data samples belong to large and dense groups, while anomalies belong to local, small and separate groups. Cluster methods apply to both supervised learning and unsupervised learning. Most methods are complex, large in size, work well with large amounts of data, and are best suited when anomalies do not form significant clusters in short time series. Examples of this type of algorithm are k-Means, Density-Based Noisy Application Spatial Clustering (DBScan), Self Organising Map (SOM), Clustering Based Dynamic Indexing Tree (CD-Tree), and, Shared Nearest Neighbour (SNN).

### 3.3.4 Similarity-based anomaly detection methods

Similarity-based methods are the most used for anomaly detection. The most common similarity-based method is the k nearest neighbours (k-NN). k-NN is a non-parametric method that requires a distance metric to measure the similarity between data observations. Euclidean distance is the most commonly used metric for data with continuous attributes. The reason for the above is that Euclidean distance does not work well for multidimensional sets, and measures such as Mahalanobis, Hamming or Chebyshev distance are used instead. The k-NN algorithm is based on a data score determined by the distance to most of the surrounding data. Therefore, new data are classified according to this assessment. However, there are some considerations to keep in mind when using this type of technique [119]:

- The lack of data can be seen as an anomaly of uncontrolled methods.
- The performance depends on the selected distance measurement method: Therefore, it is necessary to clarify the criteria when choosing a metric.
- It is valid only for low-dimensional data: Determining distance measures between instances can become more complex as data dimensions increase.

Another important similarity-based anomaly detection method is based on relative density rather than distance. This method estimates the density of neighbourhoods such that data items in less dense neighbourhoods are considered anomalous, and data items in denser neighbourhoods are considered normal. The above existing method is the Local Outlier Factor (LOF). It is based on introducing the concept of local outliers and evaluating data samples according to the average ratio of neighbourhood density to instance density [16].

### 3.3.5 Soft Computing anomaly detection methods

Soft Computing methods for anomaly detection include algorithms such as Neural Networks (NN), Fuzzy Logic, and Evolutionary Algorithms (e.g. Genetic Algorithms). These methods can tolerate uncertainty, partial data and are able to approximate predictions. A typical application of this method is in the context of security or cyber-surveillance [4], where it may be necessary to perform anomaly detection rapidly to identify an anomaly (e.g. intruder) in a system. Such algorithms can provide fast, real-time predictions for various applications [127]. Another example of the application of Soft Computing methods is in the IoT. In this case, information from sensors can be analysed to detect a possible failure, thus helping predictive maintenance.

### 3.3.6 Knowledge-based anomaly detection methods

These anomaly detection methods rely on prior knowledge about the system domain to classify whether a sample is an anomaly. They have rules that allow verification of the operating conditions. For example, a system that checks the thermal conditions of an electrical circuit, where if certain known thresholds are exceeded, the system will generate an alert indicating a possible fault or anomaly. Ruiz et al. [124] implemented a failure prognosis method based on System Operation Modes (SOM), which allows monitoring of degradation and failures in a cyber-physical system (CPS). Statistical Process Control (SPC) [88] is a method that uses 3-sigma limits (based on the variance of historical data) to determine a system's normal or abnormal operation. Figure 3.9 shows a graph with 3-sigma limits to determine the normal operation of a system.

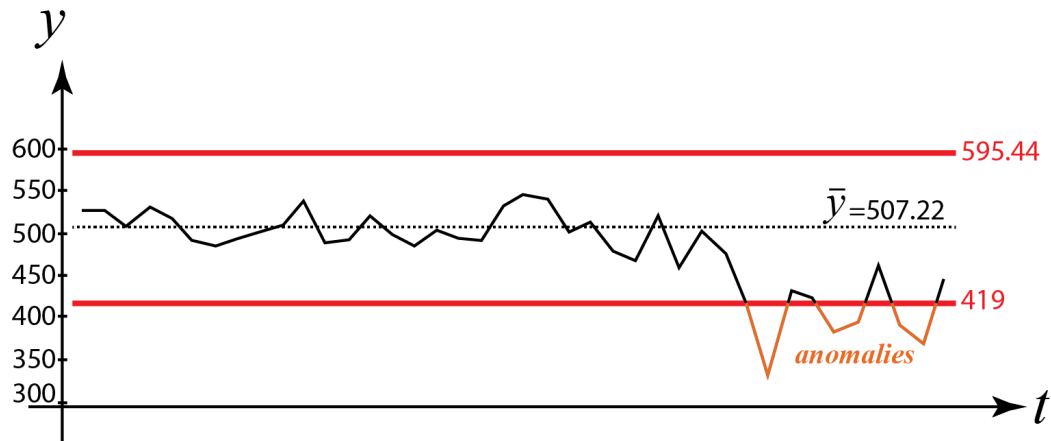


Figure 3.9: Example of SPC using 3- $\sigma$  anomaly detection.

### 3.3.7 Combination Learners anomaly detection methods

Combinations Learners are based on the ensembling of models to generate a more robust prediction of an anomaly. Simple machine learning models can present specific problems, such as statistical problems, where the hypothesis space may be too ample for the training data so that trained models may result in similar performance

metrics, and there is a risk of selecting a model that will not predict well on unseen data. Another problem that can arise is when a global optimum is not obtained. Combination Learners make it possible to exploit the potential of several models that can complement each other and thus compensate for the limitations of using each algorithm separately [148].

### 3.3.8 Time domain anomalies

Regarding time domain, anomalies can be categorised into three types: i) point anomalies, ii) contextual anomalies, and iii) collective anomalies [115, 150, 36] as seen in Figure 3.10. *Point anomalies* are characterised by a value that is a significant deviation from the expected behaviour of the data. They occur in a single period of time and in very few samples of the entire data set. On the other hand, *contextual anomalies* are samples or sequences that deviate from expected patterns in a time series of data. However, point anomalies may be within the expected values of the entire data set and are only identified in a subset of the data set. Finally, *collective anomalies* are a subset of anomalous samples concerning the rest of the data. Still, the individual samples within this subset of data may or may not be anomalous, but collectively they exhibit a suspicious pattern [28].

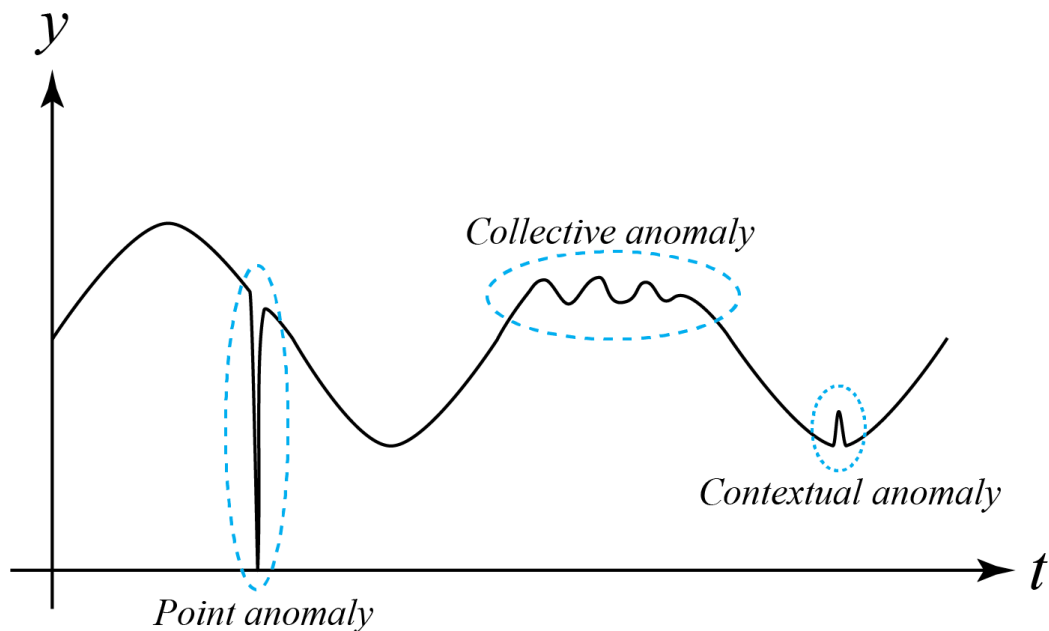


Figure 3.10: Anomaly types in a time-series plot [28].

### 3.3.9 Frequency domain anomalies

Some anomalies are easier to see in the frequency domain. Thus, transformations using Digital Signal Processing (DSP) techniques can be applied to the input data

before performing anomaly detection [119]. Fourier transform, Gabor, and Wavelets filters are examples of these transformations.

One of the methods to detect anomalies in the frequency domain is the Spectrogram. The spectrogram is a graphical representation of information about the frequency and time of a signal [19]. One way to calculate the spectrogram is first to process the Short-Time Fourier Transform (STFT) [79]. It can be calculated by dividing the signal into several data blocks using a sliding window. The Fourier transform calculates a time-dependent analogue signal in the frequency domain, but this analogue signal is typically sampled. This requires a discrete Fourier transform (DFT) to convert discrete time into the frequency domain [105, 106].

The processing required to calculate the DFT takes a long time. Computing the convolution and discrete Fourier transform requires  $N^2$  operations [34]. where  $N$  is the filter length or transform size. Using the Cooley-Tukey FFT reduces the number of operations to  $N \log_2 N$ , thus, improving computation time.

The main advantages of the FFT are computational speed and memory efficiency. DFT can be an efficient process for samples of arbitrary size ( $N$ ), but requires more computation time than FFT because intermediate results must be stored in each process, which consumes more memory [27].

When the FFT needs to be computed, the algorithm pads or truncates the input length ( $m$ ) to achieve the desired transform length ( $n$ ). The spectrogram applies this FFT to  $N - point$  data blocks to obtain the frequency content of each data block, where  $N$  is the frequency bin. STFT centres the first sliding window on the first sample of signal  $X$  and adds zeros to extend the beginning of the signal. The sliding window moves the time step samples to the following data block. When the window exits  $X$ ,  $X$  is padded with zeros. After finding the STFT, the spectrogram is calculated as the square of the magnitude of the  $STFT(X)$  elements. Figure 3.11 shows an example of anomalies in a STFT representation.



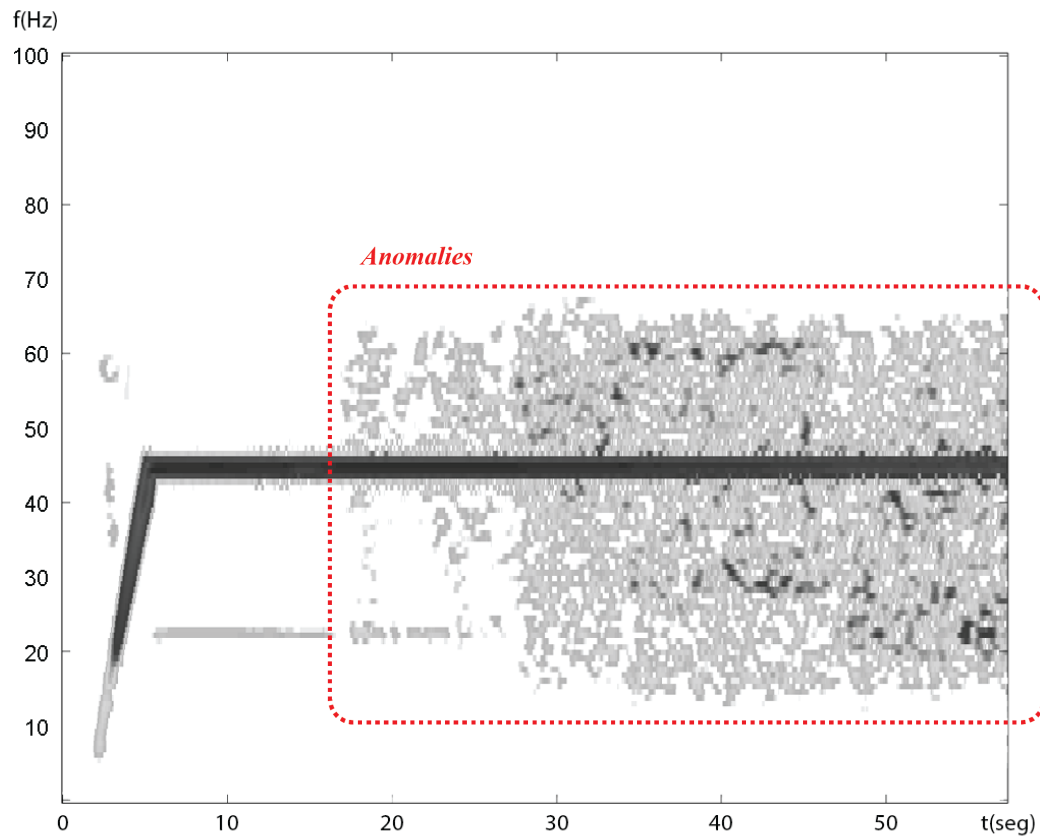


Figure 3.11: Anomalies in a STFT graphical representation [149].

The Fourth Industrial Revolution (4IR) brings new challenges to traditional industrial processes. This means either improving existing processes or creating new ones that effectively use new technologies and their full potential. In an increasingly competitive market, 4IR can be seen as a breakthrough innovation that positively impacts various industries by integrating new cutting-edge technologies [155, 78].

Common contexts for 4IR applications include manufacturing. In manufacturing, large amounts of data can be analysed in real-time to improve factory operations and production, reduce machine downtime, and ultimately improve product quality [29]. Another application context is the smart water management industry, which uses digitisation and process automation to collect helpful information from wastewater treatment plants and external sources such as weather information. AI processes this data to optimise the efficiency of the process, saving resources and the quality of the results [13, 41, 100].

However, two significant problems arise when designing a novel, AI-driven 4IR system: i) How to correctly design the system from scratch [122, 76] and ii) How to improve existing legacy systems to integrate smart-capable and connectivity layers [60]. An architecture is often used as a design baseline for building a system and it provides insights of all sub-components and connections between each of them.

## 4.1 Background

To solve the problems mentioned above, the 4IR system architecture that defines all components has been proposed by various authors at the state of the art, including technologies such as IoT, Cyber-Physical Systems (CPS), and intelligent systems. Ganti et al. [43] published an overview of mobile crowd sensing (MCS) technology commonly used for environmental, infrastructure and social applications. It presents a functional architecture showing how data from different contexts can be intercon-

nected to provide helpful information to end users.

Bagheri et al. [11] and Lee et al. [77] proposed an integration framework for incorporating CPS into production. This unified infrastructure architecture includes a five-level definition (5C Framework) that allows to develop and deploy CPS into production, from data collection to analysis and end value creation. The first layer, the “Smart Connection Layer”, receives data from machines and components using sensors from process controllers or enterprise manufacturing systems. The second layer is the “Data-to-Information Conversion Level layer”, which is responsible for extracting valuable information from the data and adding self-awareness to the 4IR CPS system. The third layer, called the “Cyber Layer”, acts as a central information hub, collecting multiple data and roughly parsing it for more information. The fourth level is called the “Cognition level”, where the 4IR CPS system generates knowledge about specific parts of the system or processes for advanced users and presents them using visual analysis tools. Finally, the fifth level is the “configuration level”. Its purpose is to provide a feedback loop from cyberspace to physical space, acting as a supervisory control that changes machines or processes depending on previous informational knowledge.

Blonda et al. [13] proposes an IoT middleware architecture and exposes its functionality as a set of cloud-supported RESTful APIs. This IoT architecture consists of three layers: i) The user layer, ii) the middleware layer, and iii) the physical layer. The middleware architecture is divided into three sub-layers: application, network, and security. According to Blonda et al., the security of an IoT system can be defined using six properties: confidentiality, integrity, availability, identification and authentication, confidentiality, and trust.

On the other hand, there is the Reference Architectural Model for Industry 4.0 (RAMI 4.0) [131], defined in three dimensions. The first dimension is the life cycle value stream, defined in standard IEC 62890, composed of the following phases: Type and Instance. The second dimension includes the hierarchy levels defined in standard IEC 62264 and IEC 61512. The levels are the following: product, field device, control device, station, work centre, enterprise, and connected world. Finally, the third dimension consists of different layers, similar to previous architectures: business, functional, information, communication, integration, and asset. Figure 4.1 provides a compact view of the architecture.

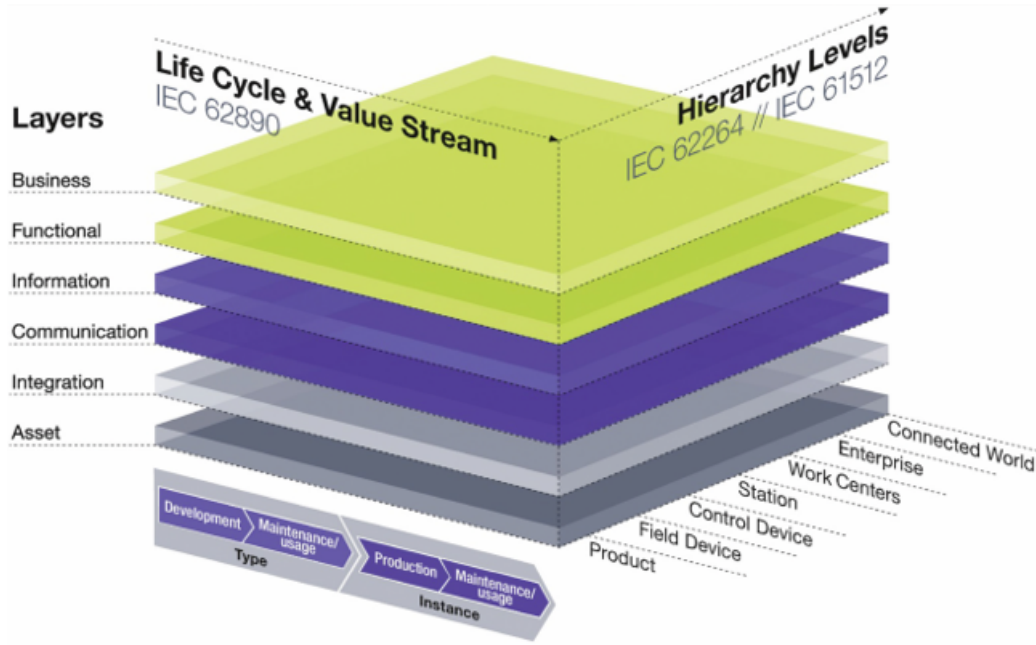


Figure 4.1: Reference Architectural Model for Industrie 4.0 (RAMI 4.0) [131].

Paiva et al. [108] propose a hybrid reference architecture called RAMI 4.0 EA that integrates RAMI 4.0 with Enterprise Architecture (EA). The proposed reference architecture allows visual and understandable enhancements to elements such as EA principles, applications, technologies, and organisational processes in RAMI 4.0, allowing companies to apply it in their 4IR projects better. Finally, the Industrial Internet Reference Architecture (IIRA) was developed by the Industrial Internet Consortium (IIC) [59]. This application is independent of the subject area, and its development was focused on the industry. IIRA focuses on industry-demanded capabilities, significantly predictive, optimisation, operations, business, analytics, and device monitoring and control.

## 4.2 Case study

SoA approaches to define the components integrated into a 4IR system have in common that they require a physical layer, middleware, and a user interaction layer. Mid-tier organisations are represented differently in different architectures, but they all emphasise the importance of security and data analytics. In addition, there needs to be guidance describing how to organise components, their connections and their interfaces in real terms to obtain a detailed system architecture and complete a working system [98]. The proposed methodology considers the most up-to-date best practices for designing 4IR system architectures from scratch and analyses case studies from various industries to understand specific problems comprehensively. Below three real industrial case studies that improved or created a new product/service by using AI-driven 4IR technologies are analysed.

### 4.2.1 Smart-Water Case Study: Industrial Wastewater Treatment Plant “La Cartuja / EDAR 4.0 Project”

EDAR 4.0 is a research project aiming to develop tools for optimising the operation and, in particular, the energy management of wastewater treatment plants (WWTPs). Different types of organisations such as water and energy engineering companies, process automation companies, WWTP equipment manufacturing companies, research centres, and universities participate in the project, all collaboratively working on different aspects of the project, finally aiming to develop a cloud-based, web platform integrating a complete set of tools for supporting an intelligent operation of WWTPs. The project’s basis consists of plant-wide data acquisition of all the processes comprising a WWTP. These processes can be classified into three principal, standard sub-processes: i) the influent process, mainly representing the input of influent water and its pre- and primary treatment, usually performed in a primary settling or sedimentation tank; ii) the biological treatment process, which is the central part of the so-called secondary treatment and represents the primary wastewater treatment process of the plant driven by different types of bacteria and protozoa, which can be complemented by additional, chemical treatments, and; iii) the effluent process, which mainly represents the output of the effluent water, either directly to the receiving waters or through a secondary settling or sedimentation tank, which is also considered as part of the secondary treatment of the plant. A tertiary treatment process consisting of additional, advanced water purification treatments for specific water uses, such as water reuse, can exist but is optional and rare. In this project, sub-processes i) to iii) of the full-scale WWTP are addressed. The processes of a WWTP in general and sub-processes i) to iii) in particular are typically controlled by one, or several Programmable Logic Controllers (PLC) integrated with different types of sensors and actuators. All the control information is then locally displayed through Human to Machine Interfaces (HMI), generally embedded within a Supervisory Control And Data Acquisition (SCADA) system. All plant information is usually shared through an industrial protocol-based Local Area Network (LAN). The above represents the basis of a typical WWTP ICT architecture. EDAR 4.0 extends this to a 4IR system architecture by setting an additional, cloud-based IoT infrastructure that can be reached through the internet. Thus, the overall WWTP ICT infrastructure must have secure access. Several services, such as multiple plants, cloud-based IoT data acquisition and storage, information monitoring (visualisation), data analysis and associated services such as Visual Analytics, plant simulation, and plant optimisation through machine learning (ML), are integrated into such a cloud-based ICT infrastructure. A specific example of accessing the above cloud IoT infrastructure and associated services could be the HTTP REST protocol. A specific example of a data analysis service could be to classify different types of water quality and predict (forecast) the evolution of water quality over time. Eventually, with the above cloud IoT platform running, WWTP data can be displayed on a web page, where water quality analyses and others can be run and monitored by remote users. Figure 4.2 details a view of the EDAR 4.0, 4IR system architecture.

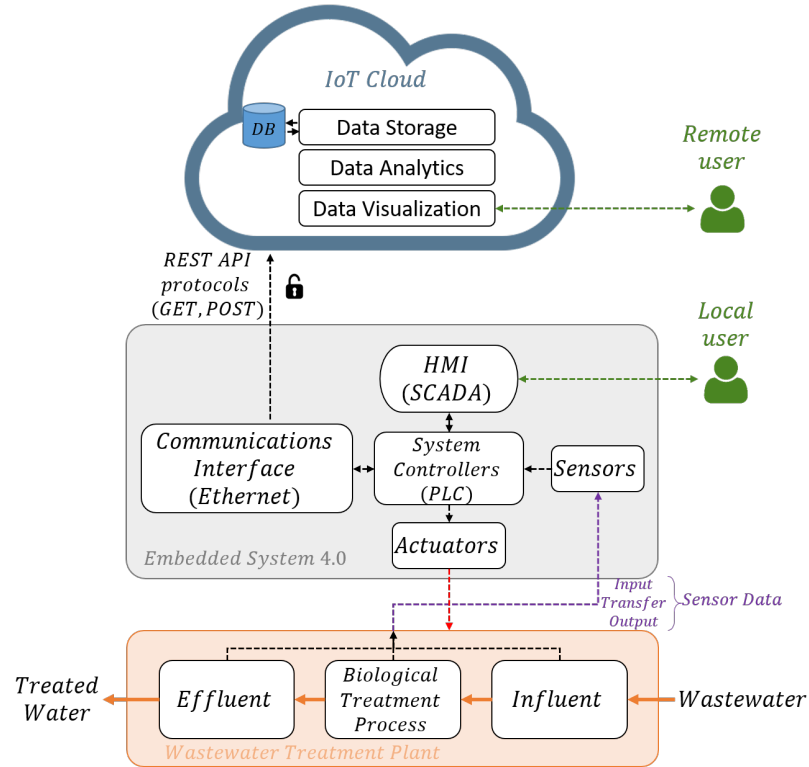


Figure 4.2: EDAR 4.0 architecture.

#### 4.2.2 Industrial Quality Testbench Case Study: Rotary Pneumatic Machines Company “MAPNER / EDAR 4.0”

MAPNER is an industrial company that manufactures rotary machines for various applications such as wastewater treatment and power generation. Once the manufacturing process of the rotary machines has finished, every machine is taken to a quality control process performed on a testbench where the machines are subjected to a set of tests in stationary working conditions in an isolated room. The outputs of the tests are then compared to some expected results described by the manufacturing order to guarantee an adequate quality and performance rate of the final product (the machine). However, that process is typically highly manual: after leaving the machine on for some time until it reaches its stationary operation region, the operator goes through a set of GUI elements of a computer program. It shows the data measured by sensors and allows manually introducing such data to a database so that subsequent calculations of physical magnitudes, such as flow rate and power, can be performed to generate a quality report of the machine. This use case is a good example of a classic manufacturing process digitalisation project where the process evolves from a view-only data management system to an automatic, real-time data acquisition and storage system, which not only improves the existing testing process (the machine’s performance can be analysed continuously instead of via a single-instant, manual data acquisition system, which can hardly reflect the overall condition of the rotary

machine) but also allows stepping forward towards a data-analysis-based machine performance study that may facilitate identifying, predicting and preventing problems for the manufactured products. The first step of adapting MAPNER's testbench to a 4IR platform was automating the acquisition of data corresponding to measurements as provided by sensors attached to the machine during the testing process and by some environmental sensors in charge of measuring relative humidity and temperature. As usual in many manufacturing processes, every sensor or measuring device has its communication protocol for providing information. Multiple protocol systems are usually managed by gateways that unify and translate the information into a standard protocol. These gateways can be independent hardware devices or software modules designed to do so. In this use case, a software gateway was implemented to gather all the information on a single, Python-based daemon (or "Python gateway") in order to be able to send the data to the data storage layer. In order to do so, the data had to be converted (unified) to a common, standard communications interface: Ethernet. Machine sensors are connected to a Siemens PLC that exposes the information through the OPC-UA protocol. The humidity and temperature information and the information provided by a set of electrical network analysers (current, voltage, power, and power factor) are exposed employing the MODBUS TCP protocol and transferred to Ethernet utilising a MODBUS-MODBUS TCP hardware gateway. The Python gateway not only made the acquisition possible, but it also helped to fix a critical aspect related to acquiring data from multiple sources - thanks to the gateway, incoming data with different sampling frequencies can be homogenised by applying data synchronisation or re-sampling algorithms prior to analysis. Once the data from the different sources is gathered and unified, the information is deployed to an internal server database, where it can be further processed and analysed to create and show operator-enriched information (such as machine status information) throughout the tests in real-time. The data storage is running within a secure LAN. In addition, a daily backup of the information managed by the data storage layer is configured. The global architecture of the entire system is shown in Figure 4.3.

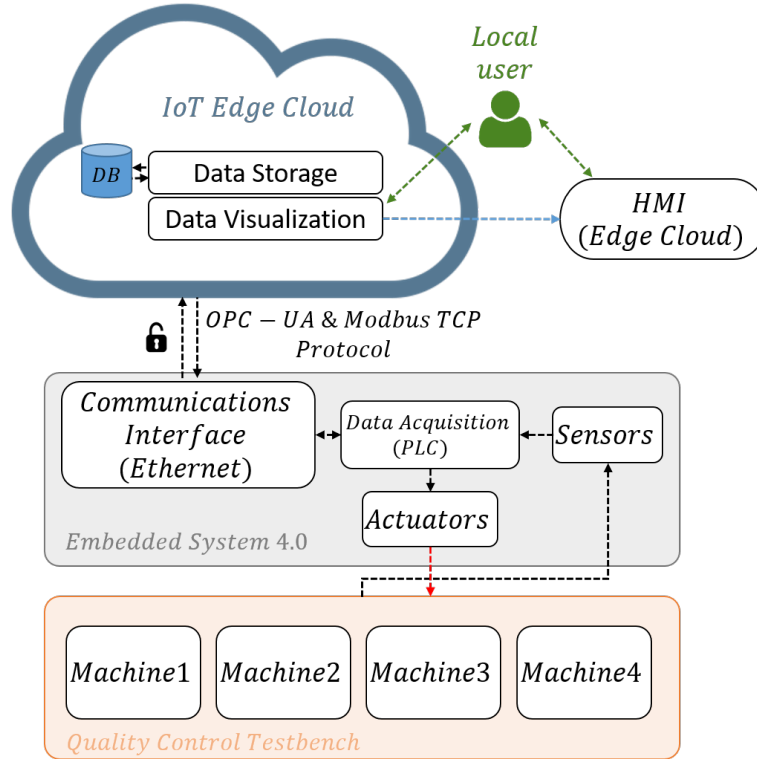


Figure 4.3: MAPNER Testbench architecture.

### 4.2.3 Smart IoT Embedded System Case Study: Rotary Pneumatic Machines Company “MAPNER / SISTELIA Project”

The third case study consists of an architecture for remotely managing data related to blower machines manufactured by MAPNER that are installed in various locations worldwide. With the arrival of the digital transformation and the 4IR, MAPNER saw the opportunity to provide their machines with greater ubiquity, especially in intelligent and predictive maintenance applications, using technologies such as IoT and AI. In this sense, MAPNER has participated in several funded R&D projects and collaborated with research agents. One of these projects is SISTELIA, which translates to “Intelligent Services for Industrial Blowers based on Digitalisation Technologies and Artificial Intelligence”. SISTELIA’s main objective is to design and implement a cloud-based data management platform based on a 4.0 embedded system called “MAPNER Panel Control” (MPC), designed to acquire, analyse, and visualise data and enriched information coming from machines that are operating worldwide, in real-time. The MPC includes an ad hoc Human Machine Interface (HMI), which gathers data from an integrated, real-time data acquisition (DAQ) system that is integrated with sensors and provides real-time information to maintenance operators both directly on the machine (local visualisation) and through the cloud (remote visualisation), via a 4G network Communications Interface. SISTELIA’s architecture consists of three parts: i) the physical blowing machine, ii) a 4IR embedded system,



and iii) a cloud data management platform. The physical blowing machine is the core process of the whole system, which operates independently of the architecture. The 4IR embedded system contains multiple sensors to gather operational data (e.g., temperature, speed, pressure, and vibrations) from the physical blowing machine, which is locally stored and processed in a real-time DAQ system. Furthermore, the locally stored and processed data is displayed to users (locally) on an integrated, tactile HMI display, where the user can program maintenance operations and configure and resolve alarms. In addition, these data are sent to an IoT cloud platform via 4G for remote storage, real-time visualisation, and data analysis. Finally, the cloud data can be monitored by remote users. This architecture is detailed in Figure 4.4.

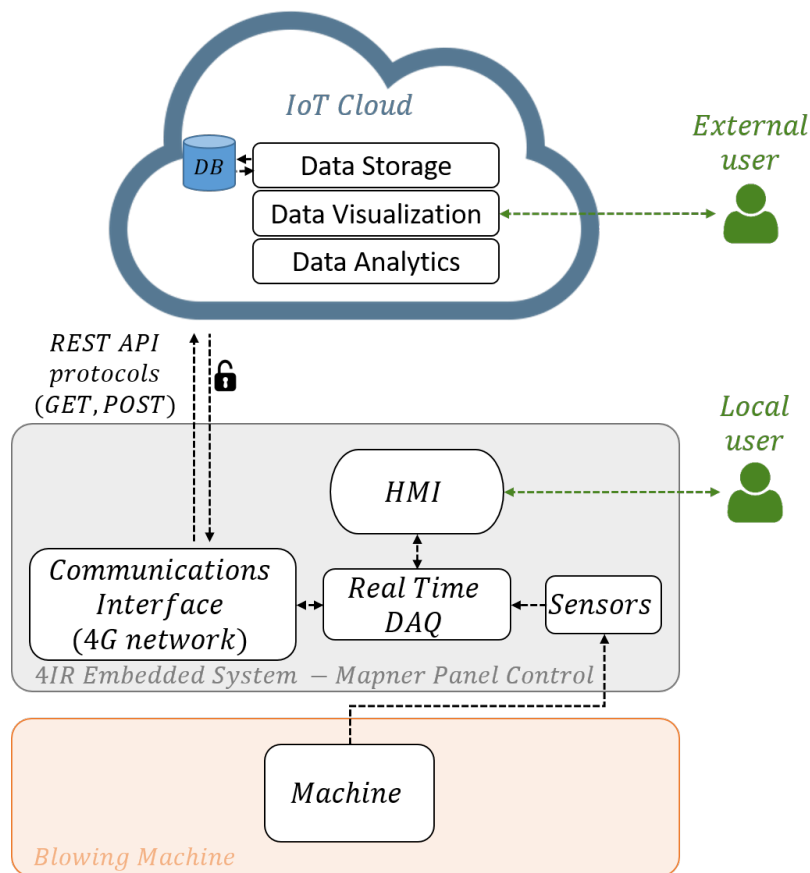


Figure 4.4: SISTELIA architecture.

### 4.3 Innovation

The main contribution of this research is a novel software-and-hardware architecture design for AI-driven Industrial 4.0 systems to facilitate the transition towards such smart-connected systems. The proposed generic 4IR architecture is intended to be used as a template for new AI-driven Industry 4.0 projects. This architecture is built around the use cases and state-of-the-art hardware and software components pre-

sented in the previous section. The proposed generic architecture presented in Figure 4.5 includes three levels: i) the physical layer, ii) the layer of an embedded 4IR system, and iii) the IoT cloud layer. The physical layer relates to the process, for example, a machine that executes a task. The 4IR embedded system comprises four (4) sublayers: The Perception and Control sublayer, where everything related to actuators and sensors can be found. The Data Acquisition and Processing sublayer, which involves the different microcontroller and PLC units with their respective internal/external storage systems for local data persistence. The Local Visualisation sublayer, which includes the different HMI interfaces for the visual and control interaction between the local user (who could be a supervisor or an operator on site) and the machine. The Communications sublayer, which addresses all the local communication interface systems such as RS232, RS485, Modbus, Profibus, and the global ones (for the new 4IR systems) through TCP/IP protocol by Ethernet, WiFi and 3G/4G networks, allowing to connect to the IoT cloud. The IoT Cloud layer incorporates four sub-layers. The first is the Security and Data Exchange sublayer, which establishes a secure connection between the Embedded 4IR System and other external information sources (External Data) through WebAPIs. The latter, for example, can use WebSockets for real-time connections, HTTP REST protocol for on-demand requests, Mosquitto Transfer Protocol (MQTT) for IoT connections, and OPC-UA for industrial data connections. The second sub-layer is related to Data Storage, where relational (e.g., SQL) and non-relational databases are used. This stored data can then be retrieved to perform different analytical and visualisation operations, for example. The third sub-layer is called Analysis, in which AI tools perform advanced processing operations. These operations create enriched information, thus adding a greater degree of knowledge about the process to allow, for example, identifying failures in a predictive manner. Finally, the fourth sub-layer includes Remote Visualisation, where dashboards are usually shown with the received data and additional graphics derived from the analysis process (e.g., Visual Analytics) for supervision by External Users.

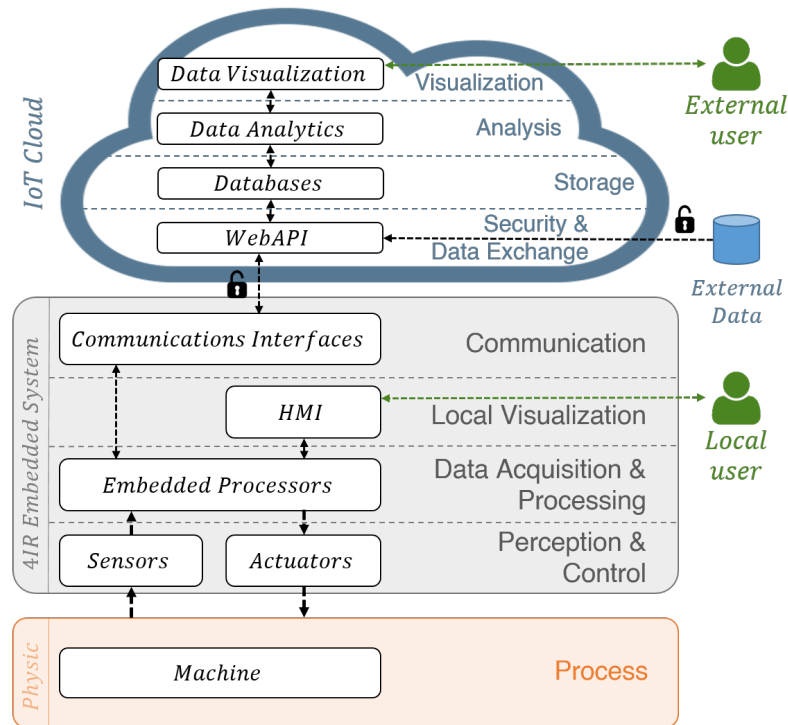


Figure 4.5: 4IR generic architecture.

## 4.4 Conclusions and future work

AI-driven Industry 4.0 systems are usually complex and challenging to understand because they comprise different hardware and software components. Commonly, these include SoA and legacy technologies. Hence, guidelines are missing explaining how such a problem could be tackled in practical terms and how components and their connections and interfaces could be organised to fully understand the system and end up with a working system. In this sense, the new hardware and software architecture for AI-driven Industry 4.0 systems has been developed based on real use cases. The architecture was created using common elements from many contemporary architectures and analysing three case studies of real industrial projects. It includes a physical layer, an embedded system layer, and an IoT cloud layer, where all components of a 4IR system can be organised clearly. This architecture provides users with a detailed view and practical guidelines for including the appropriate hardware and software components to implement a 4IR system. Future work may include methodologies for creating specific architectures based on user requirements and the above general architectures for new 4IR projects. In addition, other case studies can be analysed to expand on this architecture.

---

### Data Acquisition

---

Data acquisition is the initial part of any Machine Learning process. At this stage, it is important to consider in the design guidelines the elements that will allow the collection of information through the system's sensors and how to integrate multiple data sources to obtain more reliable information. This chapter will present a use case for the design of a cyber-physical data acquisition system for *Coffee Leaf Rust (CLR)*. Regarding phytosanitary issues concerning coffee crops, one of the main problems is the presence of pests such as the *Coffee Borer Beetle* and diseases such as *Coffee Brown Eye Spot* and the CLR [123]. In terms of disease, CLR is the most relevant disease from an economic and pathological point of view. The disease can cause massive defoliation of entire crops [101], causing devastating losses ranging from 70% to 80% of the crop in some regions of Colombia[123]. It should be noted that at the beginning of this study, the primary objective was to develop early detection of the *Coffee Brown Eye Spot* employing *Remote Sensing (RS)* and by analysing spectral reflectance data. However, after conducting interviews with Colombian experts and coffee growers, we found that this disease was not as severe as the CLR and did not restrict their economic activity. Therefore, following their indications, it was decided to integrate the *Wireless Sensor Network (WSN)* and also diagnose the CLR instead of the *Coffee Brown Eye Spot*. In this sense, the first step towards diagnosing the disease was the collection of reliable data on its occurrence. Thereby, once the necessary data had been collected, it would be possible to create a diagnostic model based on such data. Thus, this study shows two contributions: The mechatronic design of a cyber-physical data collection (acquisition) system to collect and store data, integrating RS and WSN; ii) a three-month data set for CLR detection.

## 5.1 Background

Multiple studies have been carried out, including technical methods and strategies for obtaining nutritional information, disease diagnosis and pest detection on diverse types of crops [94, 87, 140]. Recently, an important concept called *Precision Agriculture (PA)* has emerged. PA refers to agricultural management using information technology to observe, measure and respond to the variability of specific crops. PA involves applying the proper treatment method according to the plant's needs at the right time [66].

In PA, one of the methods currently used to evaluate various properties of crops is called *Remote Sensing (RS)*. RS is based on the interaction between a material and its electromagnetic radiation. This includes receiving radiation reflected from soil or plants to provide valuable information such as chlorophyll content, water stress, weed density, crop nutrients, and the presence of diseases in agricultural fields. These measurements can be made using aircraft, handheld sensors, satellites, tractors and drones [97].

Multiple authors highlight the importance of using high-quality portable devices to detect and treat diseases in hard-to-reach places. For example, Goel et al. [48] analyzed the detection of changes in the spectral response of corn (*Zea mays*) under nitrogen application and weed control. To do this, a hyperspectral sensor called the *Compact Airborne Spectrographic Imager* was employed to analyse the reflectance values of 72 bands from 409 nm to 947 nm. These bands include visible light and the emission spectrum's outer *Near-Infrared (NIR)*. Their work demonstrated the ability to detect weed infestation and nitrogen stress using hyperspectral sensors. In particular, it has been found that the optimal wavelength ranges for detection are around 498 nm and 671 nm, respectively.

In addition, Bolaños et al. [14] developed a crop classification method using the infrared and visible parts of the electromagnetic spectrum and low-cost cameras on multi-rotor aircraft. This research is based on determining a normalized vegetation index to assess health status and water content. Similarly, Chemura et al. [25] presented a method for early prediction of disease and pest infestation in coffee trees due to weak water pressure. To do this, a handheld multispectral scanner with visible and near infrared regions was deployed on an unmanned aerial vehicle. Chemura et al. also considered irrigation schemes based on specific plant water requirements.

Aside from RS, smart farming methods, and the *Internet of Things (IoT)* technologies (which refers to using intelligently connected devices and systems that use data received by sensors and actuators embedded in machines and other physical objects), there is a method called *Wireless Sensor Network (WSN)*. WSN is responsible for real-time monitoring of various agricultural characteristics. It consists of several integrated drones called sensor nodes that collect data on-site and transmit it wirelessly to a central processing station (called a base station). This station can store, process and transmit data to the Internet, where the end user can analyse the data and transform it into relevant information [10].

In this regard, Chaudary et al. [24] emphasised the importance of WSN in the field of PA using microcontroller technology called *Programmable System on Chip*

to control and define the most relevant variables in greenhouses. Their research has explored the integration of wireless sensor nodes and communication methods in the high-frequency range. This proved useful in determining the optimal irrigation strategy to meet specific crop needs. In addition, studies recommend the use of reliable low current equipment for WSN applications. In addition, Piamonte et al. [112] implemented a WSN prototype to monitor an African oil palm disease called Bud Rot. By employing sensors for humidity, pH, light and temperature, their prototype measured climate change and soil factors to determine the presence of disease-causing fungi indirectly. The study concludes that measurements of the aforementioned non-biological factors have changed slightly, which the researchers believe may reveal Bud Rot. The present state of the art shows that RS and WSN are two widely used methods in PA due to their ability to monitor various crop features and detect anomalies.

In conclusion, the state of the art shows that RS and WSN are two widely used methods within PA due to their capability to perform data acquisition tasks for identifying different crop characteristics and detect the presence of various anomalies.

## 5.2 Case study

Previous research has helped to develop a cyber-physical data collection system that can integrate both methods, RS and WSN, to diagnose CLR. By applying concepts and following best practices, a system that can collect reliable data from various sources and store it remotely can be created. Such cyber-physical systems aim to characterise coffee crops on test benches for changes caused by the disease in question. This section describes the mechatronic design of the data acquisition system.

The cyber-physical data collection system was developed following the *Pahl & Beitz* methodology [107]. The system development requirements were drawn up with the participation of the Colombian Association of Coffee Producers (*CENICAFÉ*) and the *EAFIT University*. Creating a coffee-growing data acquisition test bed that simulates various agricultural conditions and compliance with the requirements for data collection, storage and transmission were the main principles of the system design. In this sense, this section describes using the *Pahl & Beitz* methodology to create a data collection system that combines RS and WSN to diagnose CLR.

First, the *Product Design Specification* (PDS) was used to formalise, structure and classify all requirements according to their characteristics and priorities. The main requirements are to measure the physicochemical characteristics of the plant and obtain Red, Green, Blue (RGB) and multispectral images of the coffee crop on the test bench and store all this data locally and remotely. Other requirements related to the isolation and irrigation of plants, the coffee varieties used, building materials, the type of database and the protocol for communicating with field sensors.

The next step is to develop a *black box* that reflects the system's main functions. Its primary function is to collect a set of inputs, transform them, and create a set of outputs. Inputs and outputs can be divided into three main streams: matter, energy and signal, as shown in Figure 5.1. Regarding inputs, material flows consisted of CLR,

coffee plants, organic matter, fertilisers/fungicides and wind. Energy flows have been divided into electrical, human, and photovoltaic energy. The signal flow consisted of input information and expert information. The corresponding experimental coffee yield, power consumption, field sensors, and general data records were obtained at the output. This output relates to the primary goal of this study, which is to create a system that can collect, store and transmit reliable data on coffee crops of the CLR data acquisition test bed.

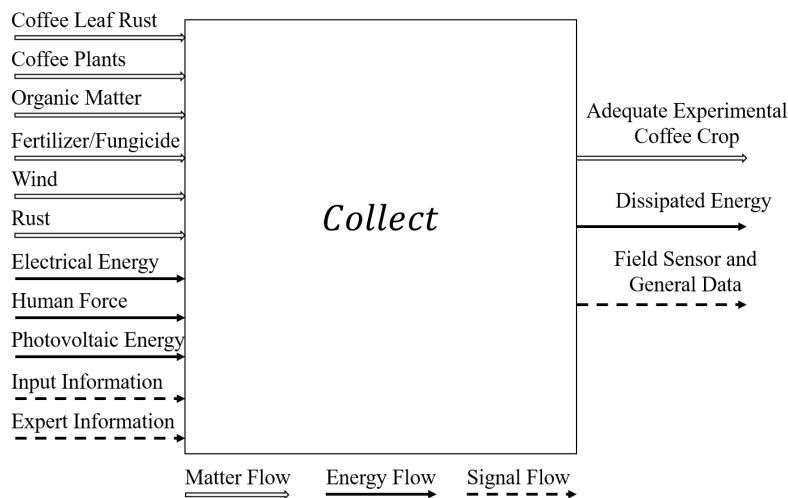


Figure 5.1: Black box representation of the cyber-physical data collection system.

After defining the black box, the functional structure was created, decomposing the inputs and outputs, and establishing a detailed understanding of the required subfunctions. For example, one of these sub-functions was to combine human power with a coffee plant to place it on a coffee plant in a test bench to affect the rear integration of a rice field sensor. In addition, sensors are used for each plot to measure soil moisture/temperature, pH, light, and ambient humidity/temperature. In addition, RGB images and multispectral images were obtained. To complete the data collection process, the data was stored locally, pre-processed for cleaning, and then sent over the Internet to a remote server. In addition, the collection process was monitored in real-time through the IoT web platform.

After determining the main features and the corresponding sub-features, a morphology matrix was built. Such a matrix presents various proposals for solutions for the implementation of each of the sub-functions presented in the function structure. The output of the morphological matrix consists of two candidate concepts, Concept 1 and Concept 2, each of which consists of different combinations of solution proposals. The concept represents two possible ways of building a data collection system, which were explored to evaluate various aspects and decide which one is most suitable for a given purpose. The resulting candidate concepts were evaluated using a scoring system that computed a weighted average of preselected evaluation criteria. These weights have been set according to the previously determined PDS and the experience of the design team. As a result, the final concept is chosen. Concept 1

was selected with 78% approval compared to 74% for concept 2, as shown in Table 5.1. The system for collecting cyber-physical data was built on the winning concept. Cyber-physical systems (CPS) are a new class of engineering systems that interact closely between cyber and physical components [70].

Table 5.1: Concept Scoring. <sup>a</sup>Value scale (score between 0 - 4); 0 = Not satisfied, 1 = Acceptable, 2 = Sufficient, 3 = Good, 4 = Totally satisfied

N <sup>o</sup>	Evaluation Criteria	Relevance (%)	Solutions <sup>a</sup>	
			Concept 1	Concept 2
1	Functionality	11	4	3
2	Simplicity	5	3	4
3	Fulfilment of requirements	10	3	2
4	Robustness	3	4	3
5	Fabrication	7	3	3
6	Assembly	6	3	2
7	Reliability	9	3	3
8	Low cost	7	3	3
9	Expert criteria	6	3	3
10	Crop management	7	3	3
11	Maintainability	3	2	3
12	Performance	8	2	3
13	Usability	5	3	3
14	Testability	3	3	2
15	Availability	10	4	4
	<b>Weighted average</b>		<b>3.13</b>	<b>2.96</b>
	<b>Total score</b>	<b>100</b>	<b>78%</b>	<b>74%</b>

The final concept is composed of a physical part and a cybernetic part. The physical part is comprised by four raised wooden beds representing the lots and separated by four plastic curtains, a rotary arm holding the multi-spectral cameras, a rain system which irrigates the lots and a circuit box with the necessary elements to interact with the electronic components. The cybernetics part of the design includes a data collector for joining the data coming from the test bench coffee-crop and a data organiser, which structures and saves it on the local storage for its posterior transfer to a remote server.

After the final concept was elaborated, the mechanical, electronic and computer design is carried out. The mechanical design considered the location of the coffee crop plots and the location of each component (e.g. electronic sensors) of the system. The electronic design considered all calculations required for each sensor or actuator of the system. In the computer design, the algorithms concerning data acquisition, conditioning and storage of the data in the IoT platforms were programmed.

Using the Pahl & Beitz methodology allowed to evolve from the definition of initial requirements to the final and detailed design of a data acquisition system integrating RS and WSN. The final 3D Computer-Aided Design (CAD) of the physical part of



the data acquisition system is shown in Figure 5.2.

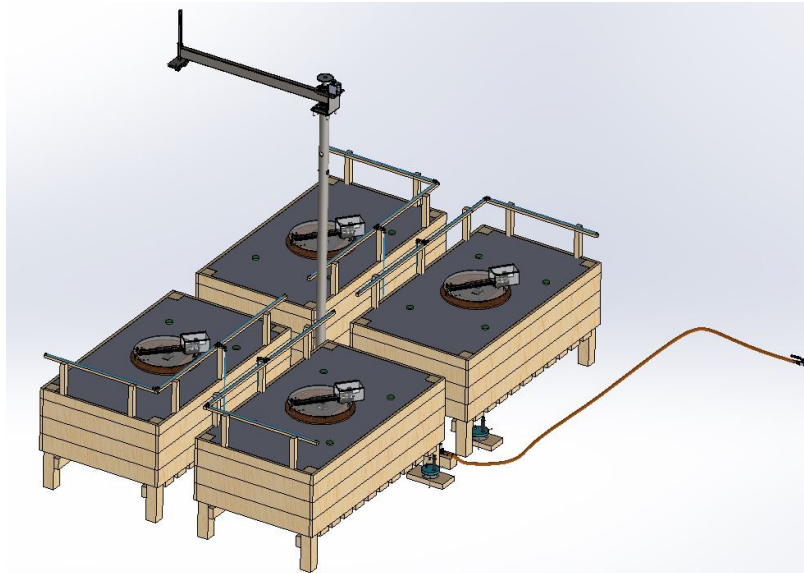


Figure 5.2: Final physical detailed design of the data acquisition system.

Figure 5.3 shows the final design of the data collection system's cybernetic part, which explains the pipeline for the execution of the data collection system.

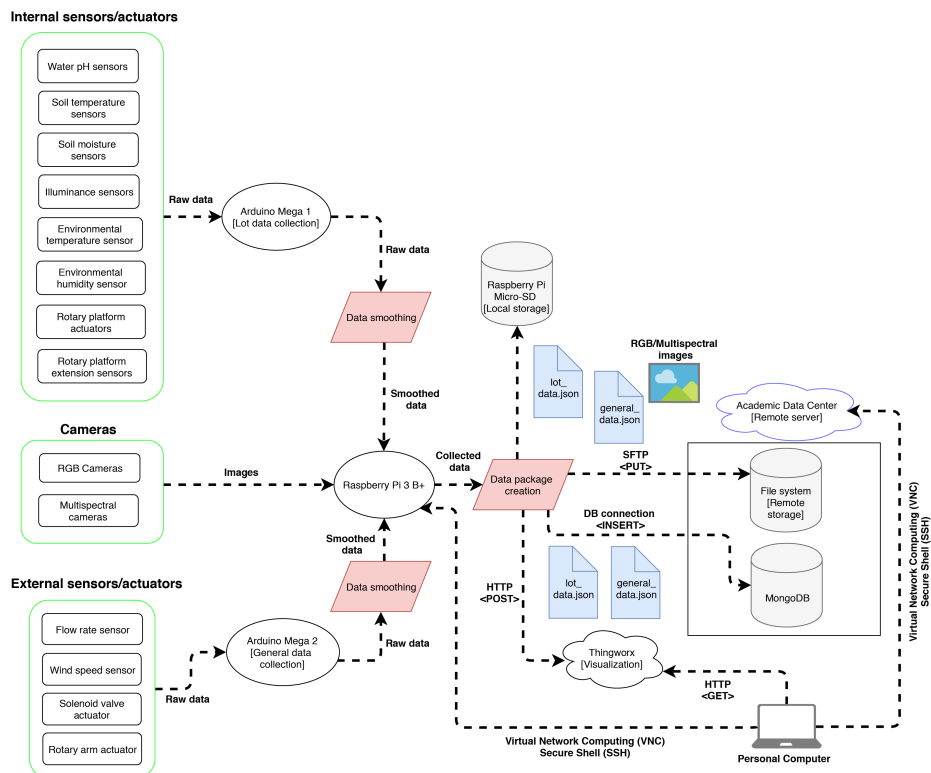


Figure 5.3: Final design of the data collection system's cybernetic part.

## 5.3 Innovation

The application of the *Pahl & Beitz* design methodology results in the target system solution. Precisely, the result of this research work corresponds to the solution obtained by applying the *Pahl & Beitz* methodology, i.e., the target system, i.e., a cyber-physical data acquisition system capable of obtaining a dataset suitable for use in the early detection of Coffee Leaf Rust. Therefore, the following describes the result obtained, i.e., the system built, which was first represented by a 3D-CAD model.

Finally, the integration of the mechanical, electronic and software parts led to the construction of a complete functional cyber-physical data collection system.

After completing the integration and construction of the data acquisition system mentioned above, a final test and calibration of each system component's operation was performed, which is essential to ensure the system's operation's reliability. For example, one part of this process included the precise adjustment of the robotic arm position with respect to each plants' lot for taking the multispectral photos, where each position was stored in the program to perform the data collection routine. Having performed the final system's calibration, a data collection routine was executed for three months. The Data Collection System recorded crop's cameras and sensors information from each lot seven times per day at different moments (with and without sunlight). It must be noted that although the data storage occurred seven times per day, the system was acquiring and monitoring the data in real-time, with a sampling period of 3 seconds. In addition to the data collected by the system, a team of biologists evaluated daily, in a separate file, the current development stage of the CLR of each data collection system lot. The output of this routine generated a dataset comprising 603 RGN files ( $\approx 153$  MB), 641 RE files ( $\approx 177$  MB), 730 RGB files ( $\approx 196$  MB) and 672 sensor data (JSON) files ( $\approx 1.12$  MB), which were ready to be used for diagnosing the CLR development stage by training a Machine Learning model.

## 5.4 Conclusions and future work

The data acquisition chapter described the mechatronic design of a cyber-physical data acquisition system that integrates RS and WSN into a coffee crop in a test bench. It can automatically collect, structure and locally & remotely store reliable multi-type data from various field sensors (pH, soil moisture/temperature, illumination, humidity/ambient temperature), RGB and multispectral cameras. In addition, a data visualisation dashboard was introduced to monitor data collection procedures in real-time. This result represents the first step towards diagnosing the CLR in *Caturra* varieties. The correct operation of the data acquisition system resulted in the creation of a 3-month data set containing the sensor and camera data needed to create the CLR design phase model. This result confirms that the developed system can collect, store and transmit reliable coffee yield data on a CLR diagnostic test bench. In future work, this data collection system may help measure and record traits that differ from

other types of crops. In addition, concerning the CLR, the data generated by this system can be used to analyse how crops respond (physicochemically and visually) to the presence of the disease. For example, artificial intelligence techniques such as computer vision and deep learning can be implemented to generate models based on collected data to diagnose the CLR effectively. The current development is intended to be a testing lab for plant experiments. However, it will be possible to conduct scalability, cost and energy analyses to turn the test lab into a complete mobile lab (using drones or ground robots) for large-scale crops. In this regard, the data set from this research work can be used to determine the optimal number and type of sensors needed in a particular case of real plantations.

---

## Supervised Ensembling

---

The ensembling of supervised models allows different sources of information and models to be used to generate an output that combines the advantages of each of these sources. This chapter will show a use case with CLR, which was already introduced in the chapter on Data Acquisition with the design of a data acquisition test bed for coffee rust. This case will present a method to combine different sources of information (WSN, RS) with DL models to diagnose the development stage of CLR, which is considered an anomaly in the coffee plantation. CLR is a fungal infectious disease that affects coffee trees and causes massive defoliation. For example, the disease has been affecting coffee trees in Colombia (the third largest coffee producer in the world) since the 1980s, causing devastating losses of 70% to 80% of the crop. Failure to detect pathogens at an early stage can lead to contamination that causes widespread destruction of plantations and seriously reduces the product's commercial value. The most common method of detecting the disease is walking around the crops and visually inspecting the plants. As a result of this problem, various studies have proven that automated technical methods can help identify these pathogens.

### 6.1 Background

Many studies have been carried out on applying technical methods and strategies for diagnosing diseases of crops [87], detecting pests [139], and obtaining nutritional information [94]. The phytosanitary status of a plantation is closely related to various essential ecosystem factors such as weather, altitude and soil type. Therefore, several biological and engineering studies aim to realise practical solutions to improve agricultural practices in maintaining healthy crops based on these factors. The most commonly used methods for effective monitoring of phytosanitary conditions include visual inspection, biological intervention, Remote Sensing (RS), Wireless Sensor Networks (WSN), and Machine Learning (ML).

Visual symptom detection uses changes in plant appearance (colour, shape, lesions, spots) as indicators of disease or pest attack [90]. In Hamuda et al.'s investigation [52], image-based plant segmentation, a process of classifying images into plants and non-plants, was used for plant disease detection [18]. For example, to assess the infection rate of the CLR in a particular lot, the number of diseased leaves of 60 random trees should be divided by the total number of leaves on those trees and multiplied by 100. Leaves are considered affected by CLR when yellow-green spots or orange dust are seen on the leaves. Depending on the number and diameter of the rust-orange spots, the severity of the disease can be divided into five categories. Figure 6.1 displays the CLR development stages.

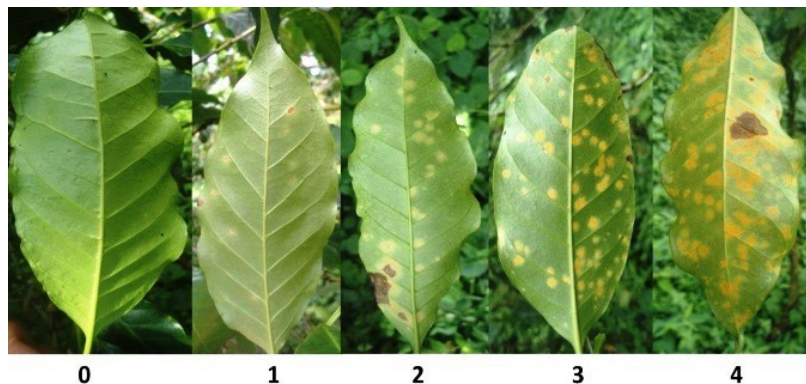


Figure 6.1: CLR development stages [31].

In terms of biological intervention, some authors note the importance of relationships between organisms living in the same environment. One of these is Haddad et al. [51], who propose a study to determine whether seven isolated bacteria selected under greenhouse conditions effectively detect and regulate CLR. To develop this study, they inoculated six *Bacillus* sp., B10, B25, B157, B175, B205 and B281, and one *Pseudomonas* sp., P286. According to the preliminary results presented by Haddad, they help detect and control the CLR in the early stages of development. Two important coffee varieties, Mundo Novo and Catuai, were chosen for the experiment due to their high susceptibility to CLR. Therefore, for three years, diseased varieties interacted with various treatments (bacteria) to analyse the evolution of behaviour among them. This was as effective as a copper fungicide in the control of diseases. Therefore, the use of biological control of the B157 isolate of *Bacillus* sp., given the harmful effects of copper-based fungicides. This can be a reliable alternative solution for managing the CLR. Thus, this study provides an opportunity to regulate CLR successfully for speciality coffee producers.

RS is based on the interaction of electromagnetic radiation with all matter. For agriculture, non-contact measurements of reflected radiation from soils and plants are used to assess various attributes such as Leaf Area Index (LAI), chlorophyll content, water stress, weed density, and crop nutrients [97]. RS helps detect problems in the agricultural sector, as it captures abnormal behaviours in crop reflectance caused by factors such as nutrient deficiencies, pests and diseases, and water stress. Calvario et

al. [17] used Unmanned Aerial Vehicles (UAVs) to monitor agave crops and integrated RS with unsupervised machine learning (k-means) to classify agave plants and weeds. Goel et al. [48] studied the detection of changes in the spectral response of corn with nitrogen application rate and weed control. To that end, the researchers employed a hyperspectral sensor called the Compact Airborne Spectrographic Imager (CASI), which has 72 wavelengths between 409 and 947 nm, which constitutes part of the visible and near-infrared (NIR) regions. The results demonstrated the feasibility of detecting weed infestation and nitrogen stress using the hyperspectral sensor CASI. Specifically, it was found that the best bands for detection were in the wavelength regions around 498 nm and 671 nm, respectively. Bolaños et al. [14] used two distinct filters, Roscolux #19 and Roscolux #2007, and an inexpensive multi-rotor aircraft camera to characterise these components using visible and infrared spectra. By this method, the anomalies that could cause crop disease could be observed and analysed. Chemura et al. [25] proposed a method for early prediction of the presence of diseases and pests on coffee trees based on invisible water stress. Their method consisted of integrating a multispectral scanner with filters in the visible, and near-infrared wavelength ranges into a UAV. Waveband scanning results showed an inflexion point between the 430 nm and 705-735 nm regions, depending on the water content of the coffee tree.

WSN is a technology used in many countries worldwide to monitor various agricultural characteristics in real time remotely. It consists of several self-contained embedded devices called sensor nodes that collect data in the field and communicate wirelessly to a central processing station known as a Base Station (BS). The BS can store, process and combine data and is responsible for sending the received data to the Internet and presenting it to the end user [10]. After the collected data is stored on a central server on the Internet, further analysis, processing and visualisation methods are applied to extract valuable information and hidden correlations to detect changes in crop characteristics. These changes can be used as indicators of phytosanitary problems such as nutrient deficiencies, pests, diseases and water scarcity. The most common sensors in agricultural WSNs are those that collect climate data, images and frequencies. Piamonte et al. [112] proposed a WSN prototype for monitoring African oil palm bud rot. This study aimed to indirectly measure climate change and soil-related factors using pH, humidity, temperature, and photometric sensors to detect pathogens.

Regarding the state-of-the-art in ML, Sulistö et al. [143] presented a computer intelligence vision sensing approach for estimating the nutrient content of wheat leaves. This approach analysed the colour features of leaf images taken in the field under various lighting conditions to estimate the nitrogen content of wheat leaves. Another work by Sulistyö et al. [142] proposed a method for determining the nitrogen content in wheat leaves using colour constancy through the fusion of neural networks and a genetic algorithm that normalises plant images at different sunlight intensities. Sulistö et al. [141] also developed a method for extracting statistical characteristics from images of wheat plants and, in particular, for estimating nitrogen content in real contextual environments where there may be fluctuations in light intensity. This work provided a robust image segmentation method using deep multilayer perceptrons to

remove complex backgrounds and fine-tune colour normalisation using genetic algorithms. After image segmentation and colour normalisation, the system's output is used as input to several standard multilayer perceptrons with different hidden layer nodes, and the simple weighted average method is used to find their output. Fuentes et al. [40] presented a robust deep-learning detector for real-time classification of various types of tomato diseases and pests. For such tasks, the detector used images from RGB cameras (multiple resolutions and different devices such as mobile phones and digital cameras). This method determined whether crops were infected with diseases and pests and, their type. Similarly, Picon et al. [114] developed an automated deep residual neural network algorithm that detects multiple plant diseases in real-time using a mobile device camera as input. The algorithm was able to detect three types of diseases in wheat crops: (i) *Septoria tritici*, (ii) tan spot (*Drechslera tritici-repentis*), and (iii) rust (*Puccinia tritici*), and *Puccinia recondita*). Chemura et al. [25] assessed the potential of the Sentinel-2 range for early detection of CLR infection levels at a devastating rate. The use of random forest (RF) and partial least squares discriminant analysis (PLS-DA) algorithms can determine such levels for initial CLR control. The researchers used the *Yellow Catuai* variety selected for its CLR susceptibility. The results of the study show that the CLR reflectance is high in the NIR region of the spectrum, as seen in leaves in the B4 (665 nm), B5 (705 nm) and B6 (740 nm) bands. These ranges achieved high overall CLR discrimination of 28.5% and 71.4% using the RF and PLS-DA algorithms, respectively. Thus, MSI-derived Sentinel-2 bands and vegetation index made it possible to detect diseases and assess CLR at an early stage, avoiding unnecessary chemical protection of healthy trees.

Several studies have been able to integrate different CLR detection tools to gain more information and accuracy in predicting this disease. In addition, determining crops' contamination level by visual inspection is a tedious task, laborious, time-consuming, and prone to human error and discrepancies. Therefore, this study assesses to what extent it is possible to diagnose the CLR stage in the Colombian *Caturra* variety by integrating RS, WSN, and DL.

## 6.2 Case study

The experimental design used in this study was completely randomised (CRD). It was used to compare two or more treatments, considering only two sources of variability: treatment and random error. This experimental design aimed to analyse whether the diagnosis of the developmental stage of CLR by integrating RS, WSN, and DL is similar to the diagnosis performed by conventional visual examination. In that sense, the study factor was the type of inspection, which had two levels ("visual inspection" and "technological integration"), and the response variable was the development stage of the disease, which was an integer number between 0 and 4. Thereby, the fundamental hypothesis to prove, presented in Equation (6.1), helped by deciding whether Treatments 1 ("visual inspection") and 2 ("technological integration") were statistically equivalent with respect to their means [117].

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \\ H_A : \mu_1 &\neq \mu_2 \end{aligned} \tag{6.1}$$

The procedure for confirming the above hypothesis was called analysis of variance (ANOVA) and required a data table with a row for each observation and a column for each treatment representing the measured value of the response variable. This procedure separated treatment variability from random error variability and compared them. If the first is higher than the second, then different means of treatment influenced the stage of CLR development, at which the type of diagnosis was determined. Otherwise, we might conclude that the means were statistically equivalent and that visual examination and technical pooling were similar to disease diagnosis. Finally, it is essential to note that the significance level used to confirm the hypothesis was 10% ( $\alpha = 0.1$ ). This is because the problem under consideration was related to agriculture, which involves many noise factors associated with changes in environmental conditions. Data collection experiments used 16 healthy, 6-month-old coffee plants from the Antioquia Gardens. These plants were kept in the EAFIT University greenhouse. A team of biologists was responsible for transplanting, farm management (weeding, fertilising, fumigation), grafting, and supervision. The biologists' team followed the process described in Chemura's study [25] for inoculation. It is important to clarify that a new group of diseased plants was kept in reserve in case the grafting of healthy plants failed over time. In addition, a group of engineers was involved in the design and assembly of a system that combines RS and WSN in one greenhouse. This allowed us to build large-scale crops, periodically record their various characteristics, store them on a remote server, and then use DL to analyse the phytosanitary situation. Thus, after the plants were inoculated and the system was tested, they were transplanted to start data collection. To do this, the large-scale crop was divided into four lots, each with specific differences in agronomic management, to replicate the different situations with actual coffee crops. This covered more scenarios and reduced false positives. Lot 1 contained four non-inoculated plants that were neither fertilised nor fumigated. Lot 2 had four non-inoculated plants that were fertilised but not fumigated. There were four inoculated plants in Lot 3, which were also fertilised but not fumigated. Four inoculated plants in Lot 4 were neither fertilised nor fumigated. Finally, the team of biologists carried out the visual inspections for diagnosis of the CLR development stage for three months. Once a day, a member of the team examined the severity of the disease for each lot and indicated the value of the response variable for each observation; this measure corresponded to the ground truth. Similarly, the technological system automatically recorded the scale crop's characteristics from each lot seven times per day at different moments (with and without sunlight because the field sensors and cameras had different illuminance requirements), assigning to each of these samples the above-mentioned daily ground truth. After the data collection phase finished, the DL diagnostic model was generated, and a comparative data table for the statistical analysis was produced based on its predictions and the results of the visual inspections. As it was expected that a considerable amount of observations would be made, only 25% of all collected data were used for the statistical study. It



should be noted too that, as was recommended, the order of the table’s entries was randomised before executing the analysis to minimise bias.

The stored data were first divided into two sets: training (cross-validated) and testing to create a suitable model for diagnosing the stages of CLR development. The training set was processed to build diagnostic models using cross-validation. After creating the diagnostic model, the test set was used to evaluate the final performance. As part of this project, a data centre was used to store remotely collected data on the physical part of the prototype. Both the MongoDB instance in it and its file system allowed copying of the SBC’s local storage, making it easy to access its information everywhere. In addition, the data centre performed data preprocessing, model building, and diagnostics during the CLR development phase. A machine learning pipeline model showing how the collected data was processed to obtain a model used to diagnose the initial stage of the disease in question is shown in Figure 6.2. This pipeline model originally consisted of four sub-directories, from Lot 1 to 4. The data will be labelled accordingly later. To this end, a team of biologists conducted visual field inspections of all plants once a day throughout the data collection phase for labelling; in this respect, they assigned an integer value from 0 to 4 was to each plant in each lot, assessed the level of damage to the leaves of the plants, and calculated the label for a particular lot as the rounded average of the disease stages of the four plants. Also, for general images, they were manually checked them one by one, leaving only those with important content (focus, brightness) and deleting the rest. A script was executed to remove extraneous sensor data files (files with missing or outlier values). The last two steps were part of the cleanup step. Finally, there are five sub-directories containing data for all lots (lot 1 - lot 4) with appropriate labelling. These sub-directories were used to create the diagnostic model and the final evaluation, given that the diagnostics were performed at the lot level.

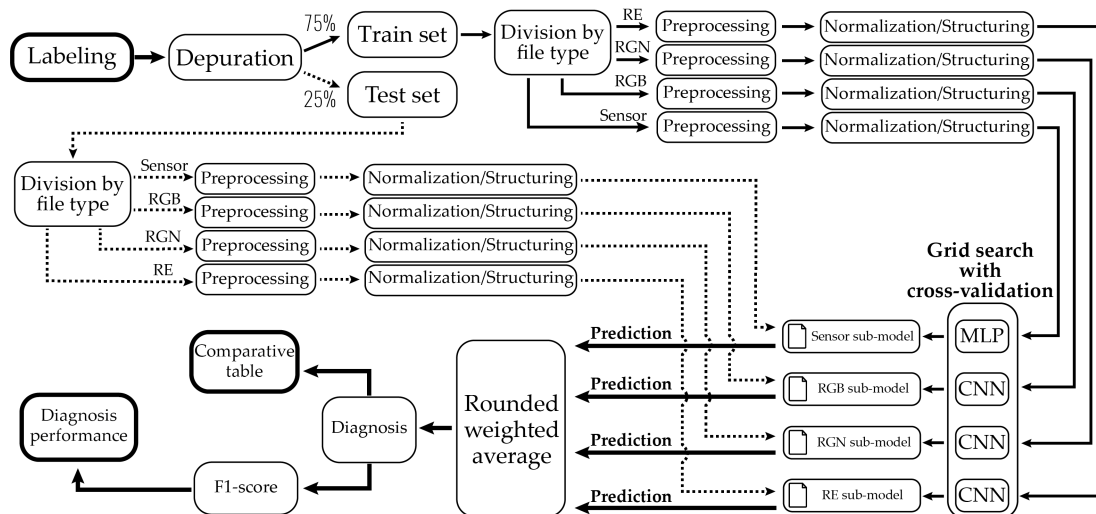


Figure 6.2: CLR ML pipeline model.

The final step consisted of combining the outcomes of the four sub-models and calculating their rounded weighted average, the weights being the respective  $F_1$ -scores.

Thereby, the definitive lot’s CLR diagnosis was obtained, and it was recorded along with the processed lot’s data directory label. Once the whole test set was covered, a table showing comparative results was generated for the statistical analysis, and the performance reached by the composite model was assessed with the calculation of the  $F_1$ -score. Table 6.1 shows the selected hyperparameters and obtained  $F_1$ -score for each of them.

Table 6.1: Hyperparameters and  $F_1$ -score for each generated submodel.

Submodel	Batch Size	Epochs	Kernel Initialiser	Activation	Rate	Optimiser	$F_1$ -Score (Cross-Val Set)
Sensor data	16	20	normal	ReLU	0.4	Adam	0.651
RGB	16	6	glorot_uniform	ReLU	0.4	Adam	0.949
RGN	32	9	glorot_uniform	elu	0.3	Adam	0.928
RE	16	6	normal	ReLU	0.4	Adam	0.878

The proposed machine learning pipeline consisted of integrating the four sub-models presented and evaluating the composite model. Throughout it, the stages of CLR development were diagnosed, a comparison table with the results achieved was created, and model’s performance calculated. For this purpose, a model evaluation script was implemented. This script loads the sub-models into memory, iterates over the test set, gets each lot data directory in it, pre-processes the contained files by breaking them down by type, and reduces the size to modify to reduce space complexity (normalise and structure each file according to the expected inputs of the sub-models and send them to the respective sub-models to get predictions. In addition, the script allowed gathering the four predicted labels and calculating their rounded weighted average, since the generated sub-models presented different performances for diagnosing the CLR development stage. Table 6.2 shows the weights for the predictions of each sub-model, which were determined as the ratio of each  $F_1$ -score in Table 6.1 with respect to the sum of all  $F_1$ -scores.

Table 6.2: Weights for the predictions of each submodel.

Submodel	Weight for Predictions (RRR table)
Sensor Data (JSON)	0.191
RGB	0.279
RGN	0.272
RE	0.258

To explain the weighted average, we assume that the sample folder containing all collected data (sensor data, RGB, RGN and RE images) was marked as CLR 2 development stage. This data in this folder was then used to develop sub-models (sensor data, RGB, RGN, and RE submodels) that produced output based on the trained model. It is also assumed that the sensor data submodel classified this as 0, the RGB submodel as 3, the RGN submodel as 2, and the RE submodel as 2. Then, given the weights from the RRR table, the average developmental stage is about 1.90. Then round up this value, and the final result of the machine learning pipeline will be *DevelopmentStage* = 2. An example of this is shown in Figure 6.3.

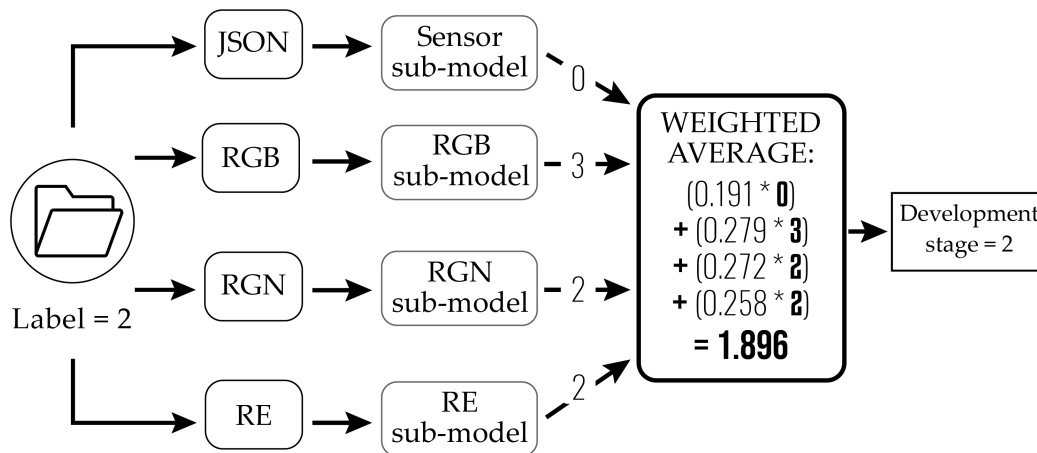


Figure 6.3: Machine learning classification example through weighted average.

### 6.3 Innovation

The result of this experiment was a composite-trained model with an  $F_1$ -score of 0.775. This model was tested using ANOVA to prove the validity of previously presented proposal hypotheses using visual inspection and technology integration methods. The resulting  $p$ -value was 0.231, greater than the significance of  $\alpha = 0.1$ . This result showed that the proposed method for automatic detection of CLR disease showed comparable performance compared to the manual/visual inspection method.

The resulting training set used to fit the submodel consisted of 968 directories. It contained 672 sensor data (JSON) files, 2192 RGB files, 603 RGN files, and 641 RE files. In addition, the test set used to evaluate the composite model consisted of 202 lot data directories containing 224 sensor data (JSON), 730 RGB files, 202 RGN files, and 202 RE files. Finally, a performance table was successfully generated after evaluating the CLR developmental stage diagnostics of the Colombian *Caturra* varieties using the DL model. Table 6.3 shows the final  $F_1$ -scores obtained by each submodel and the composite model.

Table 6.3:  $F_1$ -score reached by the individual submodels and the composite model.

Model	$F_1$ -score (Test Set)
Sensor Data (JSON)	0.570
RGB	0.920
RGN	0.946
RE	0.944
<b>Composite</b>	<b>0.775</b>

Statistical analysis of the performance evaluation results of the diagnostic model was performed using the generated dataset. The purpose of the analysis was to determine whether there was a significant difference in the mean CLR development

stage diagnosed with a visual inspection and using the proposed technological integration. This result provided the statistical support needed to answer the research question. The comparison table contained 202 observations corresponding to the diagnosed stage of development for both treatment options. Figure 6.4 shows a block diagram illustrating measurements. On the x-axis, two processes are deferred (“visual inspection” and “technical integration”), and on the y-axis, the development phase of the CLR is deferred. The visual similarity of the data distribution for each treatment suggests a possible similarity to the mean of the response variable. ANOVA was performed to assess this condition and make decisions based on hypotheses.

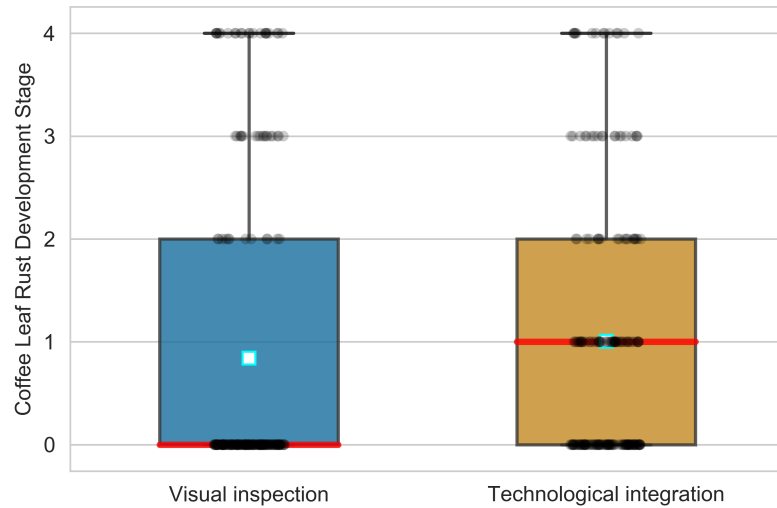


Figure 6.4: Data distribution of the observations for both treatments.

The ANOVA results are shown in Table 6.4. The p-value obtained for the treatment factor was 0.231. This value is greater than the established significance ( $\alpha = 0.1$ ), which means there was insufficient evidence to reject the null hypothesis. Thus, with 90% confidence, it was concluded that there was no statistically significant difference in the diagnosis of the developmental stage of CLR by visual inspection vs. technology integration. This result showed that both methods are significantly similar for diagnosing diseases. This study demonstrated the possibility of diagnosing the developmental stage of CLR in Colombian *Caturra* varieties by integrating RS, WSN and DL. Analysis of the results provided statistical data supporting the study's hypotheses. In this sense, the results showed the potential to complement traditional visual inspections for diagnosing the most economically limiting diseases for coffee production in Colombia, and thus technology integration could improve the phytosanitary status of coffee crops.

Table 6.4: ANOVA table of the statistical analysis.

	Df	Sum Sq	Mean Sq	F Value	<i>Pr</i> ( $>F$ )
Treatments	1	2.7	2.696	1.437	<b>0.231</b>
Residuals	402	753.9	1.875		

## 6.4 Conclusions and future work

The integration of RS, WSN and DL within this study allowed evaluating to what extent CLR developmental stages can be diagnosed in Colombian *Caturra* varieties. To this end, the most up-to-date information obtained has been synthesised, previous research works and knowledge about CLR have been detailed, and the impact of the disease on the Colombian coffee growing industry has been identified. The method was reviewed and used in the current study. A functional prototype was then created to automatically collect data in the field and transmit it over the Internet to a remote server. In addition, a diagnostic model was implemented using DL based on stored data and successfully evaluated the CLR development stage using unknown field data.

The p-value obtained from the analysis of the results was 0.231, which helped to determine with 90% confidence that visual inspection and technical integration did not show a statistically significant difference in diagnosing the developmental stage of rice field CLR. Thus, both methods of assessing the disease led to similar results, which indicates that the results confirmed the study’s hypothesis. Finally, integrating RS, WSN and DL made it possible to diagnose the CLR developmental stage of Colombia *Caturra* with an  $F_1$ -score of 0.775. This average value indicates that the diagnostic model is superior in terms of diagnostic validity and utility.

Regarding the data processing phase, a further extension of this research could include implementing a simple user interface for visualising the CLR development stage diagnosis through the generated DL model and illustrating the results to a coffee grower in a user-friendly manner. Additionally, the proposed technological integration could be scaled to a real context by using drones with one or both of the two multispectral cameras used in the experiment presented by this work as a possible approach, knowing that the identification of the CLR could be made with just one camera, e.g., RGN ( $F_1$ -score of 0.946), due to its high score. Another real context approach could be further explored using a mobile autonomous robot with a single RGB camera. Finally, the  $F_1$ -score values achieved on the test set, which showed that the submodels based on images presented a higher performance than the JSON submodel (sensor data model), suggested reconsidering the composite model for future work and focusing all efforts on improving the collection and processing of just RGB and multispectral data or using more robust sensors when the technology allows it; by using just the three submodels (RGB, RGN, and RE), we computed an average  $F_1$ -score of 0.93, which clearly showed that an improved composite  $F_1$ -score could be surely achieved, but a real context commercial application may only implement one of the best three previous submodels due to both implementation and maintenance costs.

---

## Semi-supervised Ensembling

---

In industrial machines, it is common for the data generated by their sensors not to have a label or ground truth. This is why semi-supervised ML techniques can help to train ML models with few or no labels to have enough information to create an optimal model. This chapter will present an industrial case study, where semi-supervised anomaly detection techniques and ensembling were performed to have a complete model for predictive maintenance. Most industrial companies today face problems related to system maintenance. However, some methods, including predictive or *Condition-Based Maintenance (CBM)*, can anticipate critical situations and mitigate these issues. Regarding diagnostics, preventive maintenance falls into two categories: i) models based on physical principles and ii) models based on past observations [6]. One of the methods used by the second group is the early detection of anomalous behaviour of industrial equipment [15]. This early detection avoids potential equipment failures and reduces associated maintenance costs. Anomaly detection has been explored in several application areas. Related research areas include disease detection, intrusion detection, fraud prediction, and industrial equipment failure detection. Anomaly detection typically detects anomalous conditions that do not match the normality data corresponding to the prevailing conditions. Detection of anomalous conditions is challenging, and if data needs to be processed in real-time (e.g. streaming), it poses a difficult task. Unlike batch training, where all historical data is available, and no new information is added to an already-built model, streaming training, as proposed by Silva et al. [136], has five limitations that must be considered: i) Stream data samples are received online and can be read at most once. This is a severe limitation for processing current data samples, as the system must decide whether to discard or archive them. ii) Historical data samples can only be accessed if stored in memory. Otherwise, the forgetting mechanism is used, which is responsible for discarding past samples. iii) Not all data samples can be saved, so decisions made on past samples cannot be reversed. iv) The processing time for each data sample should be short and constant. v) Data processing algorithms should produce models

that are comparable to those generated by batch algorithms. This chapter presents the evaluation and comparison of different methods to detect anomalies that, due to their performance-control metrics, establish the weight (or incidence) of each method in the final combined model, thus responding better and efficiently to the challenge of real-time anomaly detection. Specifically, the present work combines the predicted output of three Machine Learning (ML) models: Local Outlier Factor (LOF), One-Class Support Vector Machine (OCSVM), and Autoencoder employing a weighted average –using as weight the  $F_1$ -score value of each model. The goal of the combined model is the detection of anomalies in industrial systems in real-time. The proposed hybrid model was implemented using a data set from a real industrial system of air-blowing machines. Thus, it can be said that the proposed hybrid anomaly detection model applies to Industry 4.0 systems as well as other industrial frameworks where real-time data acquisition systems are available.

## 7.1 Background

Detecting anomalies in industrial environments poses two challenges. First, proposing a method to understand of heterogeneous data from different sensors (which commonly have noise). Second, obtaining an overview of normal behaviour and characterise such behaviour from historical data. Therefore, normal data behaviour must be characterised and defined to successfully detect anomalies in a dataset [118]. In addition, normal behaviour can be characterised by the following three stages. (i) Consider data describing normal behaviour through historical data (without considering anomalies) segmented into different classes according to the context in which they were recorded. (ii) Extract the most frequent behaviours, thus characterising each class. (iii) Detect anomalies in newly recorded data based on previous knowledge.

There are many studies in the literature on anomaly detection for static datasets [21]. Examples of supervised approaches are cluster methods such as SVM and Decision Trees (DT) or Distributed Matching-based Grouping Algorithm (DMGA) [26]. Other examples use self-adaptive and dynamic clustering to learn weights for anomaly detection and statistical methods such as autoregressive methods such as ARIMA models [111].

The problem with these methods is that they are not designed to handle streaming data. This is because the dataset must already be stored in the main memory. Therefore, these traditional methods are often first adapted and then applied to streaming environments. In this sense, Tan et al. [145] proposes a class of fast anomaly detections that uses only normal data and works well when abnormal data is rare. To do this, we use the Half-Space Trees (HS-Tree) algorithm. The HS-trees algorithm represents a set of random HS trees. Each HS-tree consists of a set of nodes, each of which fixes the number of data elements (called masses) in a subspace of the data flow. Mass is used to profile the extent of anomalies because it can be calculated quickly and easily compared to other methods based on distance or density. Since the tree structure is built without data, it is very efficient because the model does not need to be rebuilt on streaming data. HS-Tree only needs regular training data.

Another method worth mentioning is isolation-forest Algorithm for Streaming Data (iForestASD) [32], based on the Isolated Forest Algorithm [84]. This method processes streaming data using a sliding window. In this case, the author begins with the “concept drift”, which is a common occurrence when working with streaming data in dynamic and non-stationary environments that introduce changes in data distribution [42]. Concept drift is a problem that occurs when the statistical properties of a target variable change over time, making the anomaly detection model inconsistent with the data it processes, resulting in less accurate predictions. Therefore, the model needs to be retrained and updated to ensure effective anomaly detection based on new data received. Hulten et al. [57] proposes another anomaly detection work based on the Hoeffding Tree (HT). It is an inductive incremental decision tree algorithm used for anomaly detection. The disadvantage of this algorithm is that it requires class labels for training. Another noteworthy work was done by Laptev et al. [75]. Their system is called the Extensible Generic Anomaly Detection System (EGADS). EGADS provides accurate, flexible, scalable and extensible time series anomaly detection. This system allows us to separate forecasts, anomaly detection and alerts into three components.

Most of the approaches to detect anomalies existing in the literature are based on models that first build a profile of what is “normal” and then point out those instances that do not fit that normal profile as anomalies (statistical methods, classification-based methods, or cluster-based methods use this approach). The main objective of the semi-supervised ensembling chapter is to build an ensemble model that uses different algorithms that, by combining their results, will generate a new model to detect anomalies. Ensemble learning, either for classification or regression, refers to methods that generate multiple models combined to make a prediction [93]. Ensembles have been used in the last decades as they provide greater accuracy and increased robustness [45]. Additionally, multiple ensemble approaches have been proposed, and several studies have reported that model diversity enhances the ensemble model’s performance as different learners generalise in different ways [73].

## 7.2 Case study

The proposed ML hybrid pipeline for real-time anomaly detection consists of two stages: i) the Manufacturing stage and ii) the Operation stage. This pipeline can be seen on Figure 7.1.



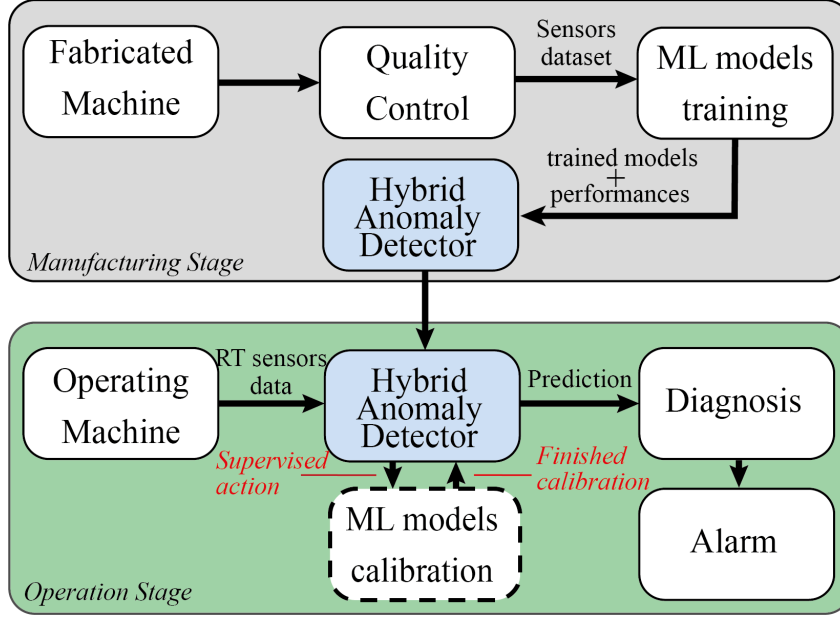


Figure 7.1: Higher-level representation of the proposed Hybrid-ML pipeline for Anomaly Detection in real-time.

The *manufacturing stage* of the proposed hybrid anomaly detection is named based on the industrial machine manufacturing process. At this stage, an ML model is trained on the machine’s quality control process data to validate that the machine meets its design criteria [67]. Therefore, the purpose of completing this manufacturing phase model-building task is two-fold: (i) to use the trained model for detecting machine design/manufacturing anomalies; (ii) to later deploy it in the operation stage of the machine when it is integrated into an industrial production process, for performing a machine operation anomaly detection task. This model construction manufacturing stage is equivalent to the design phase of a classical ML workflow. The metric chosen for measuring models’ performance is the  $F_1$ -score of label  $L$ . The data set available is a slightly imbalanced, where more machine’s “normal data” than “anomalous data” exists, for which the  $F_1$ -score metric is considered appropriate. The  $F_1$ -score is a value in the  $[0, 1]$  range, and it’s computed as the harmonic mean of the estimator precision and recall with respect to  $L$  (see Equation (7.1))

$$F_1\text{-score}_{\mathbf{L}} = \frac{2 \times \text{precision}_{\mathbf{L}} \times \text{recall}_{\mathbf{L}}}{\text{precision}_{\mathbf{L}} + \text{recall}_{\mathbf{L}}} \quad (7.1)$$

Finally, models’  $F_1$ -score ( $F_{1_i}$ ) performance ratio with respect to the sum of all  $F_1$ -scores ( $\sum_j F_{1_j}$ ) (see Equation 7.2) is calculated and used as the weight ( $w_i$ ) for the weighted average of the prediction done by each model multiplied by the computed weights. This weighted average assembles the Hybrid Anomaly Detection model at the manufacturing stage.

$$w_i = \frac{F_1\text{-score}_i}{\sum_j F_1\text{-score}_j} \quad (7.2)$$

*The operation stage* is when the machine is already running in production. In terms of traditional ML pipelines, it represents the deployment phase. Therefore, in this pipeline, machines should be able to measure the same variables obtained during the manufacturing stage via industrial sensors. Data from these sensors is captured in real-time and used as input for a Hybrid Anomaly Detector already trained during the manufacturing phase. This detector diagnoses based on the data received and generates alarms to the operator in case of anomalies. This detector can also be adjusted during operation by the monitored actions of the operator. When this action is triggered, data is captured within the time window and labelled as "normal" data. Once data capture is complete, the model is retrained within the hybrid anomaly detector. Once the calibration is complete, the system can detect anomalies in real-time.

### 7.2.1 Manufacturing-stage pipeline

The proposed pipeline requires the manufactured machines to undergo a quality control process [67]. This process allows sensors to obtain information about the manufactured equipment's behaviour over time. The data collected by the sensors during the quality control process is called the *sensor data set*.

After the sensor data is stored, the data is preprocessed for data cleansing purposes. In other words, it removes features the system cannot detect using sensors while the machine runs. The preprocessed data is then normalized so that all features are at the same scale and can be compared later in the pipeline. Then feature selection is performed to extract variables relevant to the study. This step includes domain experts as the first filter. This allows for choosing which variables to keep and which to discard. An automatic algorithm is then applied to remove redundant features [65]. Following the above, dimensionality reduction is performed using Principal Component Analysis (PCA) to extract the most representative features of the data.

The next step applies a clustering algorithm with  $k=2$  (K-means algorithm). This makes it possible to distinguish between one group of data samples belonging to the transient state and another group of data belonging to the steady state. In order to correctly label the group results obtained with the clustering algorithm, the value assigned to the cluster is first identified for the sample with the smallest timestamp in the dataset. Since this value corresponds to the *transient data group*, all samples containing the same cluster value correspond to the same state. The remaining values are labelled as the *steady-state data group*.

It is also suggested to apply the outlier detection algorithm to stationary datasets. In this case, it is proposed to use a density-based algorithm called DBSCAN. This is useful for detecting outliers in noisy applications commonly found in industrial sensor data [129].

Once the data groups relating to the transient, steady state and outliers (in the steady state) are identified, a new labelled dataset is created. In addition, a cleaning step is performed to obtain the final dataset labels. Transients and outliers are labelled with a value of -1, and stable data is labelled with a value of 1. The previous data

set is then randomly split into train, validate, and test. The training set represents 60% of all data, and only regular data is used to build each machine-learning model using cross-validation. This allows for intermediate performance testing and tuning of model hyper-parameters.

This pipeline uses the following three machine learning algorithms selected from current research on single-class anomaly detection in real-time systems. These algorithms offer the best balance between computational cost, implementation complexity, and performance. The ML algorithms selected are [2, 22, 118, 21, 20]: i) LOF, which consists of detecting anomalous data points using the local deviation of a given data point and its neighbouring data points [16]. ii) One-Class SVM (OCSVM), which detects a boundary that surrounds most of the data (ordinary data) and the following new data beyond the boundary is considered anomalous [128, 30]. iii) Autoencoders, which reduce the dimensions of input data by encoding information into a smaller space. This compressed space is decoded to the same dimensions as the original input. Then, the reconstruction errors in this process define possible anomalies [5].

Normal data are used for the training because the proposed pipeline is designed to identify anomalies based on a single class for novelty detection, and individual ML models use unsupervised algorithms.

The validation set, which corresponds to 20% of the data set, is used to obtain the definitive performance (in this case, the  $F_1$ -score value) of each trained model. The weights for the predictions of each model are then determined as the ratio of each  $F_1$ -score value (obtained using the validation set). The weights are stored to be later used for the rounded weighted average of the Hybrid Anomaly Detector component. The test set corresponds to the final 20% of the data set and is reserved for measuring the performance of the hybrid anomaly detector. The manufacturing stage pipeline is shown in Figure 7.2.

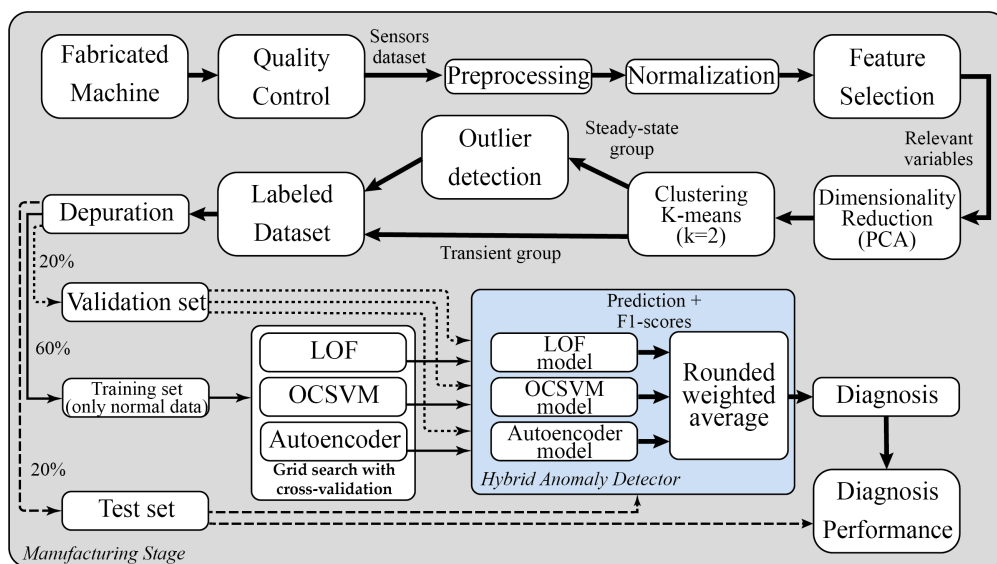


Figure 7.2: ML Manufacturing stage pipeline.

### 7.2.2 Operation-stage pipeline

This step is performed while the machine is running. During this process, the operating machine generates real-time data from pre-installed sensors that match the same sensors used during the production (machine manufacturing) phase. Each run cycle is pre-processed and delivered to the previously captured hybrid model to diagnose if the machine is healthy or if anomalies should be reported via alarms. The operation stage also allows for calibrating the Hybrid Anomaly Detection models required in industrial systems that degrade over time and can be planned (for example, at every maintenance). The operator must verify that the machine is in a stable state and under optimal conditions of normality and activate the ML models' calibration routine to carry out this process. Once this process is activated, the system will collect data during a period of time, which will depend on each system's dynamics. Each data will be stored with the normality label in the data set. This data set with normal data is then used to retrain each ML algorithm with cross-validation. Finally, the newly trained models are updated in the Hybrid Anomaly Detector. It should be noted that only the weights (obtained through the  $F_1$ -scores) that were acquired in the manufacturing process are used because, in the operation process, usually, there are no anomalous data to measure this performance. The operation stage pipeline can be seen in Figure 7.3.

Operation Stage pipeline

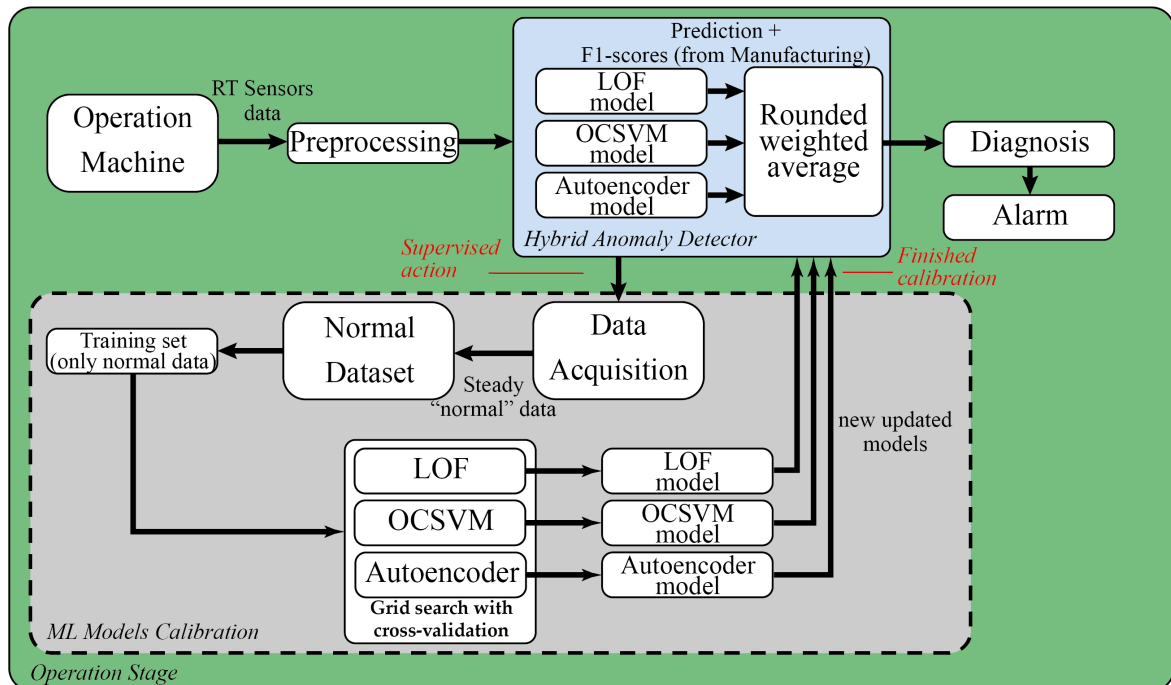


Figure 7.3: ML Manufacturing stage pipeline.

The proposed machine learning real-time anomaly detection hybrid pipeline has been tested on three different industrial air blowers from a local industry using quality control datasets, and these machines are currently in operation. The machine data

collection period is from January 7, 2020 to October 2, 2020. Data is recorded and saved at 2-second intervals. The final dataset consists of 16 columns (15 variables and timestamps) containing 1990 observations for machine A, 2009 observations for machine B, and 2132 observations for machine C. These dataset characteristics are shown in Table 7.1.

Table 7.1: Air-Blowing Machines' data set characteristics.

<b>Model version</b>		
A	Date control	1 June 2020
	Start-End time	08:30 - 09:36
	Total Samples	1990
	Normal Samples	67.789%
	Anomaly Samples	32.211%
	Sample period	2 sec.
B	Date control	15 June 2020
	Start-End time	08:13 - 09:20
	Total Samples	2009
	Normal Samples	55.351%
	Anomaly Samples	44.649%
	Sample period	2 sec.
C	Date control	14 July 2020
	Start-End time	09:40 - 10:51
	Total Samples	2132
	Normal Samples	70.779%
	Anomaly Samples	29.221%
	Sample period	2 sec.

The sensor data set consisted of the variables measured by the sensors attached to each machine during the quality control step. Measured variables are flow, power, water temperature, nozzle temperature, inlet pressure, outlet pressure, flow temperature, machine vibration, rpm, active power, cos phi, motor current, motor voltage, ambient humidity, ambient temperature environment and atmospheric pressure. Common variables are selected for the production and operation phases in the preprocessing phase. Variable preprocessing can be seen in Table 7.2. A total of 11 variables (both production and operational) were selected. In addition, samples with invalid or missing values were checked in the preprocessing step and removed from the dataset.

Table 7.2: Variables preprocessing at Manufacturing Stage.

Variable	Available at Manufacturing	Available at Operation
Flow Rate	✓	×
Nozzle Temperature	✓	×
Suction Pressure	✓	✓
Discharge Pressure	✓	✓
Flow Temperature	✓	✓
Machine Vibrations	✓	✓
RPM	✓	✓
Active Power	✓	✓
Cos Phi	✓	✓
Motor Current	✓	✓
Motor Voltage	✓	✓
Ambient Humidity	✓	✓
Ambient Temperature	✓	✓
Atmospheric Pressure	✓	×
Water Temperature	✓	×

The last step of the proposed ML pipeline consisted of implementing an ensemble of three models: LOF, OCSVM, and Autoencoder, through a weighted average distribution. Table 7.3 shows the weights for the predictions of each model, which were determined as the ratio of each  $F_1$ -score value in the validation set with respect to the sum of all  $F_1$ -score values for each class (“-1” and “1”). As an illustrative example, for a given sample, the LOF model predicted an anomaly (-1), the OCSVM predicted normality (1), and the Autoencoder predicted an anomaly (-1) again, each output is multiplied by its respective weight, this computing the final classification of the hybrid model. Thus, considering the weights from Table 10, the output of the hybrid model will be 0.8. If this value is positive (greater than 0), the hybrid model will classify it as a normal data point (“1”), whereas, if is negative or zero (lower or equal than 0) it is considered as an anomaly.

Table 7.3: Weights for the predictions of each submodel.

Model	Machine	Weights (-1)	Weights (1)
LOF	A	0.373	0.363
	B	0.412	0.378
	C	0.215	0.259
OCSVM	A	0.406	0.371
	B	0.417	0.370
	C	0.177	0.259
Autoencoder	A	0.352	0.359
	B	0.344	0.353
	C	0.304	0.288

## 7.3 Innovation

In addition to the proposed pipeline for real-time anomaly detection, the proposed hybrid model should represent improved performance measures of individual models. In this case, the accuracy, recall, F1 scores, and area under the ROC curve (AUC) of all models were compared.

### 7.3.1 Manufacturing pipeline results

Three machines were selected corresponding to three different model versions to confirm that the hybrid model performs equally well on different hardware. The confusion matrix allows seeing what types of hits and errors (type I or false negative errors and type II or false positive errors) the current model have through various metrics such as accuracy, reproducibility, sensitivity, and specificity. The ensemble model confusion matrix was analysed to see if individual models perform better. In this regard, we will focus on two indicators. i) Accuracy: Abnormal data is classified as normal. Also called false positive rate (FP) or Type I error. ii) Recall: Normal data is classified as abnormal, also known as false negative rate (FN) or type II error. The Confusion matrix for machine A, machine B, and machine C are shown in Tables 7.4, 7.5, and 7.6 respectively.

Table 7.4: Machine A - Confusion Matrix (Test Set).

<b>Model LOF</b>		<b>Predicted</b>	
Actual	Anomaly (-1)	120	39
	Normal (1)	2	237
		Anomaly (-1)	Normal (1)
<b>Model OCSVM</b>		<b>Predicted</b>	
Actual	Anomaly (-1)	130	29
	Normal (1)	4	235
		Anomaly (-1)	Normal (1)
<b>Model Autoencoder</b>		<b>Predicted</b>	
Actual	Anomaly (-1)	63	96
	Normal (1)	97	142
		Anomaly (-1)	Normal (1)
<b>Model Hybrid</b>		<b>Predicted</b>	
Actual	Anomaly (-1)	132	27
	Normal (1)	1	238
		Anomaly (-1)	Normal (1)

Table 7.5: Machine B - Confusion Matrix (Test Set).

<b>Model LOF</b>		Predicted	
Actual	Anomaly (-1)	122	31
	Normal (1)	3	271
		Anomaly (-1)	Normal (1)
<b>Model OCSVM</b>		Predicted	
Actual	Anomaly (-1)	137	16
	Normal (1)	23	251
		Anomaly (-1)	Normal (1)
<b>Model Autoencoder</b>		Predicted	
Actual	Anomaly (-1)	56	97
	Normal (1)	98	176
		Anomaly (-1)	Normal (1)
<b>Model Hybrid</b>		Predicted	
Actual	Anomaly (-1)	126	27
	Normal (1)	4	270
		Anomaly (-1)	Normal (1)

Table 7.6: Machine C - Confusion Matrix (Test Set).

<b>Model LOF</b>		Predicted	
Actual	Anomaly (-1)	174	46
	Normal (1)	2	180
		Anomaly (-1)	Normal (1)
<b>Model OCSVM</b>		Predicted	
Actual	Anomaly (-1)	163	57
	Normal (1)	5	177
		Anomaly (-1)	Normal (1)
<b>Model Autoencoder</b>		Predicted	
Actual	Anomaly (-1)	170	50
	Normal (1)	51	131
		Anomaly (-1)	Normal (1)
<b>Model Hybrid</b>		Predicted	
Actual	Anomaly (-1)	177	43
	Normal (1)	2	180
		Anomaly (-1)	Normal (1)

The confusion matrix shows a generalised improvement of the hybrid model's performance compared to the other models in all three machines, both for recall and precision. For the experiments being analysed, precision should be maximised as much as possible since it is indicative of the anomalous values detected by the system.

Tables 7.7, 7.8, and 7.9 show the models' summary results, both individually and jointly, using their metrics for comparison.



Table 7.7: Machine A - Metrics table (Test Set).

Model	Label	Precision	Recall	$F_1$ -score	AUC
LOF	-1	0.980	0.750	0.854	0.873
	1	0.860	0.990	0.920	
OCSVM	-1	0.970	0.820	0.887	0.900
	1	0.890	0.980	0.934	
Autoencoder	-1	0.390	0.400	0.394	0.495
	1	0.600	0.590	0.595	
Hybrid	-1	0.990	0.830	0.904	0.913
	1	0.900	1.000	0.944	

Table 7.8: Machine B - Metrics table (Test Set).

Model	Label	Precision	Recall	$F_1$ -score	AUC
LOF	-1	0.980	0.800	0.877	0.893
	1	0.900	0.990	0.941	
OCSVM	-1	0.860	0.900	0.875	0.905
	1	0.940	0.920	0.928	
Autoencoder	-1	0.360	0.370	0.365	0.504
	1	0.640	0.640	0.643	
Hybrid	-1	0.970	0.820	0.890	0.905
	1	0.910	0.990	0.946	

Table 7.9: Machine C - Metrics table (Test Set).

Model	Label	Precision	Recall	$F_1$ -score	AUC
LOF	-1	0.990	0.790	0.878	0.890
	1	0.800	0.990	0.882	
OCSVM	-1	0.970	0.740	0.840	0.856
	1	0.760	0.970	0.851	
Autoencoder	-1	0.770	0.770	0.771	0.746
	1	0.720	0.720	0.722	
Hybrid	-1	0.990	0.800	0.887	0.897
	1	0.810	0.990	0.889	

From the table above it can be seen that the hybrid model's performance improves the individual models' performance. Thus, it justifies combining models with hybrid models using weighted averaging to improve the final performance of the entire pipeline. Also, note that the results presented by the autoencoder are relatively low compared to other models. This is because autoencoders are well suited for anomaly detection using time windows and convolutional network architectures, which is not the case. The problem with a convolutional architecture is that it requires time

windows that could add significant delay in the operation stage and would make it difficult to compare its metrics to those of the rest of the models due to the transformation of the training, validation, and testing data that is needed to be done for being able to use the data with this type of model.

### 7.3.2 Operation pipeline results

The anomaly detection algorithms described above are ineffective if the trained model cannot be processed smoothly in a standard real-time operating environment. To measure performance, data batches containing 2012 samples were compared against all individual models on a regular computer (8 GB RAM and minimum Intel Core i5 or equivalent, no graphics card required). The computation time required to obtain the result was measured. Finally, the same data was used with the hybrid model, and the computation time required to process the data was logged. The results are shown in Table 7.10.

Table 7.10: Performance results of each model in microseconds.

	<b>LOF</b>	<b>OCSVM</b>	<b>Autoencoder</b>	<b>Hybrid</b>
<b>mean</b>	803.6	175.4	34445.7	35982.6
<b>std</b>	2515.4	21.3	9999.4	11254.8
<b>min</b>	674.6	159.8	30300.3	31399.4
<b>max</b>	112896.7	446.4	174986.8	187873

As expected, the hybrid model was slower than the individual models. However, its response time still exceeds the real-time response threshold defined for mainstream computers in 2020 (less than 200ms for the worst-case cycle for batch analysis), allowing real-time anomaly detection.

## 7.4 Conclusions and future work

This chapter has developed and presented a Hybrid Machine-Learning Ensemble for Anomaly Detection for a Real-Time Industry 4.0 System (employing semi-supervised ML algorithms). This ensemble consists of implementing two stages inspired by a standard industrial system: i) A Manufacturing Stage and ii) An Operation Stage. Up to our knowledge, there are no other ML methods that consider these industrial stages. The ensemble system was tested on three machines, presenting an increased  $F_1$ -score value and AUC concerning individual ML submodels (LOF, OCSVM, and Autoencoder). The ensemble model for Machine A presented a  $F_1$ -score value of 0.904 for anomalies (-1), a  $F_1$ -score value of 0.944 for normal data (1), and an AUC value of 0.913; the ensemble model for Machine B presented a  $F_1$ -score value of 0.890 for anomalies (-1), a  $F_1$ -score value of 0.946 for normal data (1), and an AUC value of 0.905; finally, the ensemble model for Machine C presented a  $F_1$ -score value of 0.887 for anomalies (-1), a  $F_1$ -score value of 0.889 for normal data (1), and an AUC value of 0.897.

The proposed system allows vertical scaling in the number of algorithms used for the ensemble. As seen in section Results, subsection B, the hybrid model presented a maximum computation time of approximately 190 milliseconds, fast enough for real-time anomaly detection. Concerning individual models' performance, the Autoencoder results showed a low  $F_1$ -score value, so it is proposed to test other algorithms (e.g., Isolation Forest, Elliptic Envelope) to improve the overall performance of the whole assembly. However, a study of the computational cost linked to the retraining of more types of algorithms must be carried out.

Future work is proposed to study system retraining in the Operation Stage pipeline and its computational cost. It is also proposed to study the proposed system developed on machines with different levels of degradation. Additionally, a data imputation study should be carried out to generate synthetic samples for systems where some information is missing (a loss of data due to communication breakdowns is a common problem in industrial systems). Deep Learning techniques could be considered when creating meta-classifiers using different base classifiers such as recurrent neural networks, like LSTMs, where time series need to be considered. Furthermore, a study with a larger number of machines must be carried out to see how well the hybrid model generalises against the individual sub-models. In cases where the hybrid model does not provide any improvement, other ensemble strategies such as taking the best of the individual sub-models are considered.

Finally, as this project focuses on single-type anomaly detection, a challenge to be addressed in future work will be to be able to classify or categorise different types of faults. For that, the authors might use appropriate methods such as explainable ML or correspondingly labelled datasets.

---

## Frequency-based Anomaly Detection

---

In industrial systems, there can be anomalies that are not easily visible in the time domain. Some of these anomalies can be identified in the frequency domain. In this chapter, a case of anomaly detection will be presented for detecting cracks in transport of hygroscopic particulate compressed material using frequency transformations. The transport of goods has been carefully studied, as it is essential to ensure the quality of the final product. The particulate matter situation is further complicated when companies decide to innovate product geometry due to the trade-off between packaging and cargo space optimisation. It refers to particulate compressed hygroscopic materials that can be solved by compressing particles into geometric shapes to improve the end-user experience. However, if the transported material is compacted particles, the problem is that cracks and damage to the product may occur if the truck operating conditions, such as vehicle suspension and road conditions, are not met during transportation. This chapter will introduce a crack identification method applied to hygroscopic particulate compressed materials subject to simulated transport conditions. An experimental approach is used to simulate package and transport conditions. Spectral analysis was used to determine if a material fulfils transport requirements to go from a given location to its destination, in terms of cracking.

### 8.1 Background

In the simulation of transportation situations using vibration, frequency analysis can be used to obtain the necessary information. One of these types of analysis is the spectrogram. Spectrograms can be used to analyse sound patterns (such as animal sounds) [104], radar and sonar applications for target tracking [37], and medical applications such as measuring blood flow using ultrasound information [146] and detecting cracks in cantilever beams [80]. Using accelerometers as sensors to measure vibration, Gillich and Praisach [46] proposed a method based on changes in nat-

ural frequencies for detecting damage in beam structures, concluding that damage changes the natural frequencies. Sha et al. [132] presented a new method to detect single and multiple damages in beams using relative natural frequency changes, enabling damage in cracked beams to be identified and measured. Onchis [103] also used the frequency spectrum to identify cantilever damage using the Gabor transform and his proposed procedure with LASSO minimisation. Sinou [137] investigated the possibility of detecting the presence of open cracks in rotating machinery at low and high rotor accelerations. Webb [152] measured for the first time the full spectral response of a Fiber Bragg Grating (FBG) sensor exposed to vibration. Yan [156] used a multi-scale enveloping spectrogram through vibration signal analysis for the health diagnosis of bearings, Puchalski [116] used vibration signal to diagnose mechanical defects, Wang [151] extracted fault features with transient vibration signal analysis. Jweeg et al. [68] performed a frequency analysis from the vibration of the pipe to investigate the effects of cracks in the pipe, and it was found that the more profound the crack, the lower the frequency. Aramburo-Londoño et al. [7] presented a dynamic analysis using the Finite Element Method (FEM) to evaluate the effects of vibration on hygroscopic particulate matter. The results infer the behaviour of compacted powders in handling and transport and determine ideal conditions for product packaging. Gomes et al. [49] proposed an experimental approach to validate the canonical Power Spectral Density (PSD) by acquiring acceleration data from an electrodynamic shaker and proposed software for signal processing. Wu et al. [153] presented a method for detecting and locating fatigue cracks in aluminium plates by measuring instantaneous baselines using a series of piezoceramic transducers and a shaker testbed. Aymerich et al. [9] investigated the effect of boundary conditions on nonlinear acoustics that can be used for impact damage detection of composite structures. In addition to health monitoring applications, Shin investigated the two properties of correlation coefficients between two transient vibration signals used for Location Template Matching (LTM) methods that can estimate the impact location through vibration signal analysis. Most of the methods presented in the literature are based on damage to solid materials but not particulate compacted materials. This chapter presents a methodological proposal on combining different frequency domain analyses to deal with vibrations in particulate compression materials.

## 8.2 Case study

The proposed method is integrated into the normal compaction and transport process and usually starts with product compaction at certain speeds ( $P_C$  and  $S_C$  respectively). The compressed product is then packaged in a special envelope designed to protect the product from damage. The packaged products are then stacked in the vehicle's storage, and pressure ( $P_P$ ) is applied to the packaged product below, causing it to vibrate at a frequency of ( $f_T$ ). This frequency is mainly dependent on road conditions and vehicle suspension. After the vehicle reaches its destination, the product quality is verified, and the defective products are rejected. Finally, the process is repeated. The proposed method tests a product sample to see if it will withstand

transport conditions, which are simulated through hardware and software. The hardware consists of a vibration test bench that first vibrates the sample with simulated spring pressure ( $P_P$ ) at a specific frequency ( $f_T$ ). Sample vibration data is collected using accelerometers through a data acquisition device, and the information is transferred to a computer database. The software component is a developed spectrogram post-processing algorithm that checks the stored database and indicates whether the sample failed the test (cracks/anomalies were detected) or passed the conditions of simulated transportation. Figure 8.1 shows the contribution of this research work and how this new method relates to conventional compaction and transport.

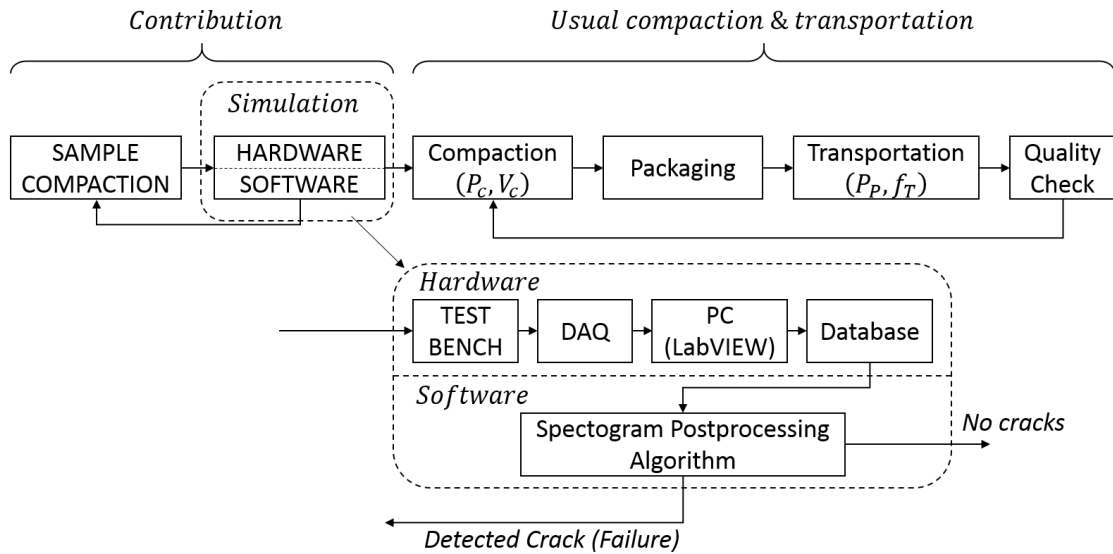


Figure 8.1: Proposed new method for crack detection.

A series of practical experiments were conducted to test the proposed method, based on frequency analysis, for detecting cracks in fragile compressed materials during transportation. The test consisted of a simulated packaging and shipping testbed that applied preload and vibration to a compacted hygroscopic particulate material to see if the material would crack. Based on typical load handling conditions, a vibration frequency of 45 Hz is programmed, and the resulting acceleration is obtained from data acquisition. In this test, five hygroscopic materials were used to validate the proposed method, including powdered sugar, plaster, white cement, chocolate powdered drink, and orange powdered drink. In addition, the results allow materials to be compared concerning their behaviour under specific transport and packaging conditions. The software tool used to measure crack initiation was a spectrogram. The expected result of this method is the detection of loss of uniformity through the detection of new frequency components. These components appear when the cracked parts vibrate along with the sample. This behaviour means that the sample did not survive the transportation conditions, as cracks began to appear on the spectrogram.

For analysing the acquired vibrations' data, an algorithm was developed using spectrogram function plus a direct-form FIR low pass digital filter [138] for data

conditioning and filtering possible electronic noise for frequencies over 100 Hz. The inputs for the low pass digital filter are shown in Table 8.1 and the output coefficients were used in the integrated function *filter*.

Table 8.1: Low pass digital filter inputs.

Variable	Value
FIR Design Method	Equiripple
$F_s$	50000 Hz
$F_{pass}$	90 Hz
$F_{stop}$	100 Hz
$A_{pass}$	0.1 dB
$A_{stop}$	40 dB

The result of the algorithm is an image file containing a graphical representation of the colour map of the spectrogram. The x-axis is the time in seconds, the y-axis is the frequency, and the black intensity is the amplitude of the vibration signal. An example of a graph of results for a compressed sample subjected to vibration on a test stand is shown in Figure 8.2. Four states can be seen: i) vibration-free start, ii) steady state vibration frequency, iii) initial segmentation (first crack) of the sample and iv) total crack. The first crack can be detected when a new frequency and oscillation frequency appear.

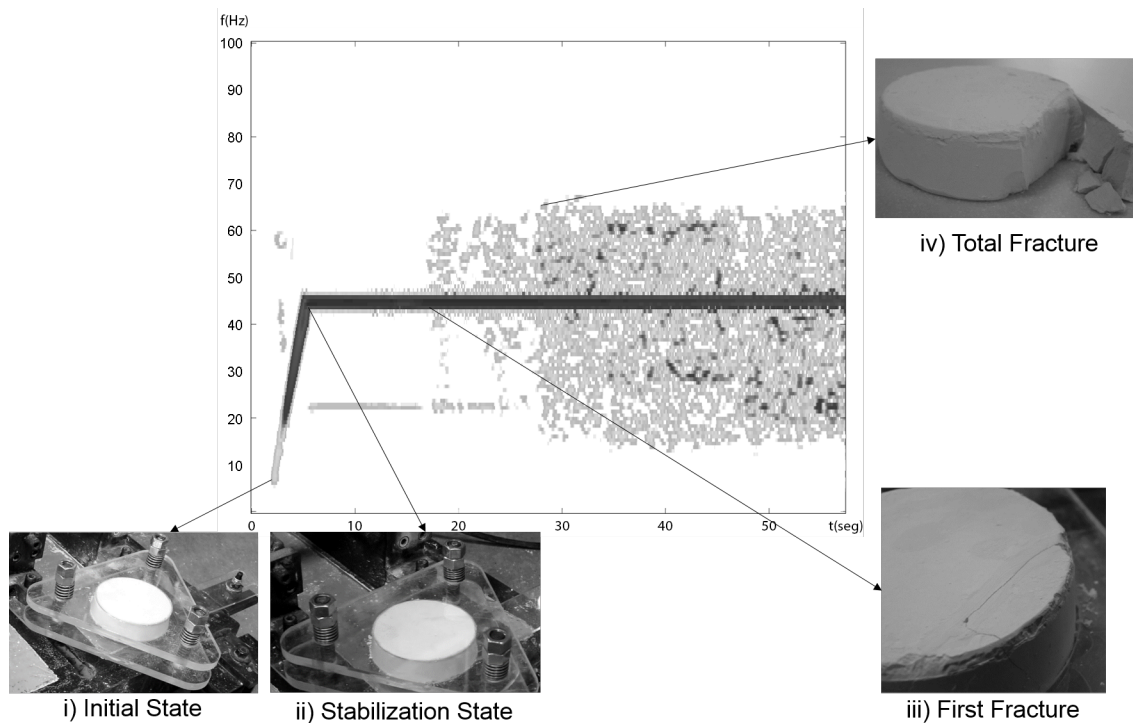


Figure 8.2: Sample result spectrogram with the corresponding states.

## 8.3 Innovation

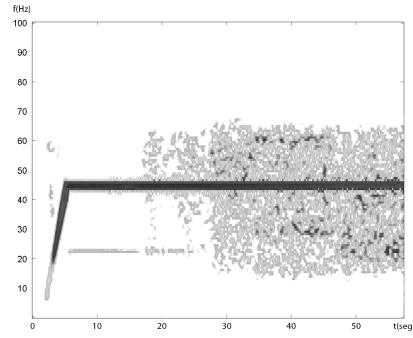
Four samples of different hygroscopic particulate materials were compacted with a pressure of 110 PSI, each of them having a different particle size in order to validate the proposed method. These particulate materials were:

1. White cement
2. Chocolate powdered drink
3. Orange powdered drink
4. Plaster

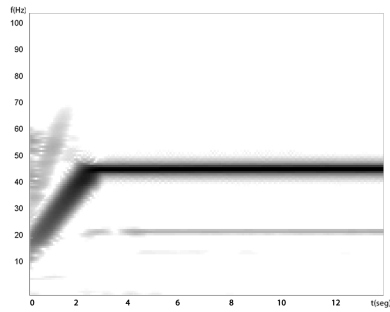
After compaction process, each sample was tested through the vibrations testbed at 45 Hz. Then the acquired data was processed using the test processing method in order to check if they will fail or not with a spring load of 3 Kg (package load simulation) and a vibration frequency of 45 Hz (truck frequency during transportation).

After the test, it was found that only two samples got a crack: the Sample-1 and Sample-3. This deduction was made using the method proposed in the section 3, where crack occurs when more components in frequency appear than the main frequency of vibration (45 Hz). The Sample-1 showed the first crack at  $t = 19$  seconds and at  $t = 29$  seconds it had a total crack. This result concludes that the compressed white cement under the given conditions will have cracks during transportation. Sample-2 did not present even a minor crack, which means this compressed material can be transported without getting a crack using the previous conditions. Sample-3 had the first crack approximately at  $t = 11$  seconds and after  $t = 20$  minutes, it did not present a major change which means that this type of mixture (Chocolate and Plaster) will present a minor crack during transportation under these conditions. Sample-4 showed similar results to Sample-2, it did not present even a minor crack, which means this compressed material can be transported without getting a crack using the previous conditions. Figure 8.3 shows the spectrogram results.

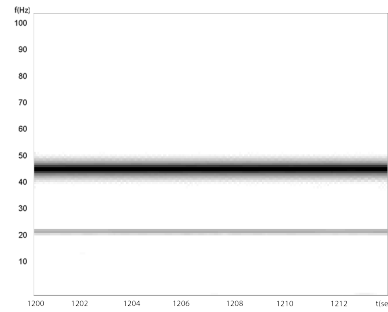




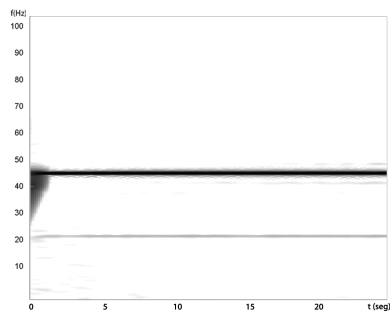
(a) Sample-1 - Crack detection.



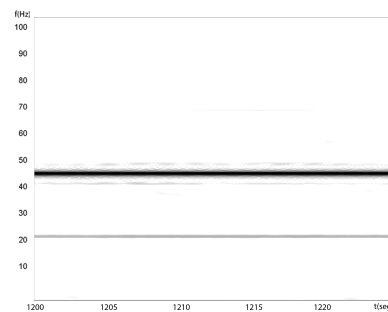
(b) Sample-2 - First 14 secs.



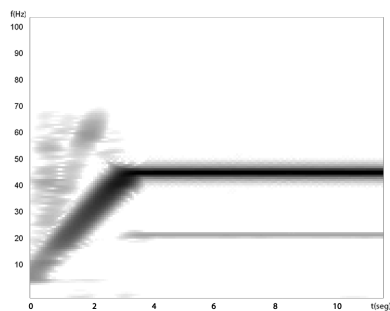
(c) Sample-2 - After 20 mins.



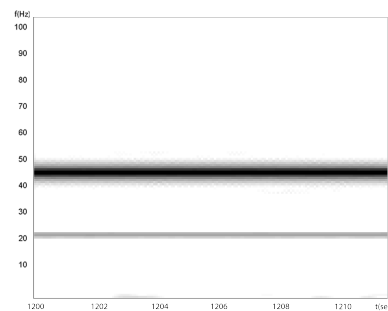
(d) Sample-3 - After 25 secs.



(e) Sample-3 - After 20 mins.



(f) Sample-4 - After 12 secs.



(g) Sample-4 - After 20 mins.

Figure 8.3: Spectrogram results for all samples.

## 8.4 Conclusions and future work

Frequency analysis was used to check if cracks occurred during compacted products' transportation to detect product failure during transportation. This method considers all possible interactions between vehicle vibrations and vertical loads acting on the sample (stack of the same product) from the compacted material, the package type and the compacted product. Cracks (anomalies) in compacted hygroscopic compressed materials can be detected by finding the time on the spectrogram at which new frequency components appear, different from the fundamental vibration frequency. When these new frequency components are detected, they indicate that particles that have left the main sample have begun to oscillate around the material. Of the experiments presented, the most stable materials under test conditions were sample 2 (chocolate powder) and sample 4 (a mixture of plaster and orange powder), indicating that these materials are safe to transport compared to the other sample materials under standard freight transport conditions. This test is generic and applicable to other types of compressed materials. This analysis checks the integrity of these materials during shipping and checks the compression conditions during the design of the compression product, which reduces the risk of cracking during shipping, and the quality of the final product after shipping can be improved. It has been proven that the amount of data received increases linearly with the sampling time to obtain reliable results. This requires significant computing power to analyse the spectrogram. Secondly, it is recommended to use parallel computing to reduce simulation time. Future work may improve the algorithm for determining the exact failure time. Vision acquisition can also be implemented on packaging and transportation simulation test benches to compare spectrogram and vision results to make testing more reliable. For this study, the round shape of compacted samples was considered, but tests of other compacted sample shapes were further analysed to determine the handling stability of hygroscopic compacted materials.



---

## Visual Analytics

---

The last stage of the Machine Learning for Industry 4.0 process is Visual Analytics (VA). The output of the models created and the information gathered must be clearly visualised for the end user so that they can make more objective decisions based on the data. This chapter will present the case of a Waste Water Treatment Plant (WWTP), where a visual analytics-based platform for WWTP that allows users to determine relationships between data through simple data validation is developed.

New connected industrial facilities are rapidly generating data that needs to be stored, processed and monitored in real-time to make decisions that optimise production in new Industry 4.0 factories. This newly generated data and methods for visualising it present several challenges, including dimensionality reduction and real-time visualisation of high-dimensional data. A data processing and visualisation solution is visual analytics. Keim et al. [69] defined visual analytics as a combination of interactive visualisation and automated analysis techniques for better understanding, reasoning, and decision-making based on vast and complex datasets. Visual analytics focuses on creating new tools that allow users to: i) synthesise information that allows getting new insights from massive heterogeneous sets of data, ii) detect current states of systems and discover possible new states, iii) provide real-time assessments and perform actions based on these assessments.

Keim et al. [69] also proposed six visual analytics challenges: i) scalability with large data volumes and dimensionality, ii) graphical representation of data quality, iii) visual representation of the level of detail, iv) new display interface walls such as massive power, v) visual analytics score frameworks, vi) real-time update interactions. Many of these issues still need to be addressed today.

The success of a WWTP can be managed by finding optimal process conditions and identifying factors, features or patterns critical to data-driven decision-making. Newhart et al. [99] emphasised that wastewater treatment plant operators usually keep a fair amount of historical data. Additionally, recent advances in data-driven process control and performance analysis, as well as greater computing power, could

allow the wastewater industry to lower costs and improve operations, as well as the required experience of data cleansing and processing specialists in the field of data processing, limits the possibilities for getting the most out of data.

One of the most critical factors influencing decision-making in the age of big data is finding relevant data and extracting meaningful insights from it. To address this issue in the context of WWTP, the Estación Depuradora de Aguas Residuales 4.0 (EDAR 4.0) project has developed a suite of WWTP management and operation systems, integrating cloud computing, data analytics, and visual analytics. The goal of EDAR 4.0 is to improve the storage, processing, computing and decision-making capabilities for wastewater plant operations [89]. The results of EDAR 4.0 have been tested and validated at La Cartuja (Zaragoza, Spain), a complete municipal WWTP operated by the Girona-Veolia company.

The five variables analysed for the operation and management of wastewater treatment plants in EDAR 4.0 are Biological Oxygen Demand-5 ( $BOD_5$ ), Total Chemical Oxygen Demand ( $TCOD$ ), Total Kjeldahl Nitrogen ( $TKN$ ), Total Phosphorus ( $TP$ ) and Total Suspended Solids ( $TSS$ ). These variables were chosen based on the European Directive 91/271/EEC and are quality requirements that wastewater from WWTPs must comply with. Similarly, where applicable, wastewater treatment plants are located in areas declared sensitive to eutrophication (Aragon, Spain), so specific values for total phosphorus and total nitrogen are required. Table 9.1 shows the quality requirements based on the above European directives.

Table 9.1: Water Quality requirements from European Directive 91/271/EEC

Variable	Absolute Values	Performances
$BOD_5$	25 $mgO_2/L$	70%
$TCOD$	125 $mgO_2/L$	75%
$TKN$	10 $mg/L$	90%
$TP$	1 $mg/L$	80%
$TSS$	35 $mg/L$	70%

## 9.1 Background

The most common approach to optimising process performance for fluctuations in incoming water quality is to apply process control and process simulation to obtain optimal operating strategies. Ordinary Differential Equations (ODEs) are widely used in process modelling. To model a WWTP using an ODE, it is important to first model the steady state of the process under a given set of disturbances and operating conditions. The disadvantage, however, is the longer computation time when parsing the ODE. Recently, Jong-Rack et al. [63] proposed an improved Newton-Raphson method to reduce computation time. The above shows that there is still active research on modelling wastewater treatment plants using ODE.

In a separate study, Flores-Alsina et al. [39] developed a plant-wide water phase chemistry model that describes pH change in conjunction with industry-standard

models. Flores-Alsina et al. formulated the general equilibrium as a set of differential-algebraic equations (DAEs) instead of ODEs to increase simulation speed. In addition, Flores-Alsina et al. applied a multivariate version of the Newton-Raphson algorithm to handle multiple algebraic interdependencies.

It is essential to mention that the International Water Association (IWA) reference simulation model has been available for several years, providing a platform for comparative analysis of activated sludge management strategies. Jeppsson et al. [61] extended the IWA benchmark to facilitate the development and evaluation of the effectiveness of plant-level control strategies, thus including both processes that consider wastewater pre-treatment and sludge treatment.

Finally, although Li et al.'s [83] work is not related to wastewater treatment, it is a combination of ODE and ML, so it is worth mentioning. Their paper presents a neural Fourier operator for turbulence simulation with zero superresolution. This work showed high speed and accuracy compared to conventional solvers.

The ODE approach to process control and modelling has several drawbacks. Large ODEs perform significantly worse than database models, such as visual analysis of complex data. Wastewater management solutions exist, but some are difficult to use and understand for those who need to learn the details of such solutions.

Visual analytics is a way to visualise data and simplify decision-making. It combines interactive visualisation with data analytics and machine learning (ML) to help people analyse, explore and understand data at any scale. The visual analysis process can be summarized using the scheme proposed by Van Wijk [147] (see Figure 9.1). The first step is extracting the data from the data stored in the database or the data stream. This data is then analysed and processed to extract the most important features presented in the visualisation phase. The render step then creates an image to represent this processed and selected data, or is created to a user specification. The user then sees and recognises this image, gaining information and knowledge from the latest images. This step is repeated until the user has scanned the entire image. Finally, users can generate hypotheses. Hypotheses are developed at the stages of research and analysis. In addition, a new analysis may be required, translating into specification steps where users can interact with current visualisations to gain new knowledge.

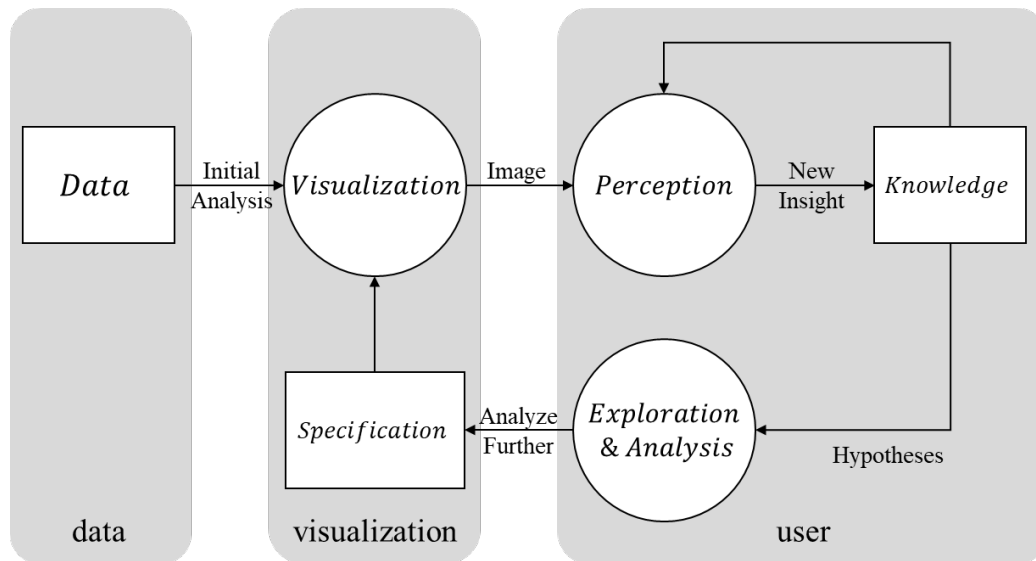


Figure 9.1: Visual analytics process framework [147].

As Liu et al. [85] mentioned, interactive model analysis, the process of understanding, diagnosing, and refining machine learning models using interactive visualisations, allows users to solve real-life artificial intelligence and data mining problems effectively. In Liu et al.'s article, work related to visual analytics is classified into three categories: (i) understanding, (ii) diagnostics, and (iii) refinement. Liu et al. point out that many methods create static images to show which parts of the image are most important for classification. However, interactive visualisations are essential in understanding and analysing models, helping people understand different machine-learning models better. Therefore, our proposal is about the dynamic creation of demand-driven models, such as water quality models and how their responses can help us understand specific variables.

Visual analytics is widely used in industrial contexts. Jonker et al. [64] used a visual analytical approach to understand better complex time series models applied to economic data. Using computational linguistics, visual analytics and deep learning techniques, Chang et al. [23] analysed hotel reviews and responses collected on TripAdvisor to determine response strategies. Park et al. [110] proposed a visual analytics system to improve supply chain managers' decision-making process. Sun et al. [144] created PlanningVis, which consists of a visual analysis system that supports the exploration and comparison of production plans with three levels of detail: a plan overview that shows general differences in the plan, a product view that visualises the various characteristics of individual products, and a production detail view that shows products. Finally, Wu et al. [154] reported on developing and implementing an interactive visual analytics system. The system allows shop floor managers and operators to use domain knowledge and apply significant human decisions to drive automated analytical approaches to produce understandable and trustworthy results in real-world applications.

In WWTP, visual analytics tools allow to quickly and interactively explore mul-

tiple views of the same multidimensional data. It is possible to have a global view of data behaviour through different colours, orientations and data. Interactive visualisation of trade-offs across multiple dimensions is suitable for stakeholders with different interests [91]. Kim et al. [72] recently introduced the Operator Decision Support System (ODSS) to help wastewater plant operators make the right decisions. Kim et al. describe a system of fluctuations in water quality in wastewater treatment plants. Kim et al.'s system consist of two diagnostic modules, three prediction modules, and a scenario-based help module. The prediction module is based on the k-nearest neighbours (k-NN) method and predicts water quality three days in advance. Wastewater treatment plants account for an increasing share of operating costs associated with electricity consumption. Piao et al. [113] used mathematical modelling to propose six improvements to reduce power consumption. The power consumption of the proposed Piao et al's plan was estimated using an artificial neural network.

## 9.2 Case study

The methodology followed in this article is inspired by the proposal of [8], who argued that these are the typical steps in a successful data analysis and mining:

1. Data collection and acquisition. It is the process of gathering and measuring information on targeted variables; it is divided into the following activities:
  - (a) Analysis of data origin and frequency.
  - (b) Quantification of data uncertainty.
  - (c) Compilation of data from various sources.
2. Data management and data validation. It consists of checking the accuracy and quality of source data before using, importing, or otherwise processing it. It is compound by the following activities:
  - (a) Definition of erroneous data.
  - (b) Detection and removal of outliers based on the variable analysis.
  - (c) Detection of outliers based on physical processes.
3. Data visualisation. It is the graphical representation of information and data, its main activities are:
  - (a) Exploration and visualisation of data.
  - (b) Development of intuitive, powerful visuals.
  - (c) Development of algorithms for prediction of future conditions.

AvRuskin et al. states that "due to the physical nature of waste-water process data, it is recommended that laboratory, operations, and engineering staff be consulted at all points in the process to confirm assumptions" [8].



By using the above methodology (see chapter 'Architecture'), an EDAR 4.0 architecture is created. This architecture has the WWTP process as the base, which is a factory-level data acquisition of all the processes that make up a WWTP. This process can be divided into three main standard sub-processes. First, the inflow represents the entry of incoming water and its preliminary and primary treatment, usually in a primary settling tank or settling tank. Secondly, the biological treatment process is central to the so-called secondary treatment. It is the primary wastewater treatment process for various types of bacteria and protozoa using chemicals. Third, the wastewater treatment process is a product of the wastewater treatment plant. This output receives either directly treated water or water that passes through a secondary settling tank or settling tank that is considered part of the plant's secondary treatment. Typically, a wastewater treatment plant's processes and sub-processes are controlled by one or more Programmable Logic Controllers (PLCs) that combine various sensors and actuators. All control information is displayed locally via a Human Machine Interface (HMI), usually integrated into a Supervisory Control and Data Acquisition (SCADA) system. Based on industrial protocols, all system information is usually transmitted over a local area network (LAN).

EDAR 4.0 expands this to a 4IR system architecture, creating an additional cloud-based IoT infrastructure accessible via the Internet, which requires (secure) access to the entire WWTP and ICT infrastructure. There are various services in this cloud, such as multiple WWTPs monitoring systems, 4IR data cloud collection and storage, information monitoring (visualisation), data analysis, visual analysis, factory simulation, and related services, such as factory optimisation through automatic learning. A specific example of access to the above cloud IoT infrastructure and related services can be the HTTP REST protocol. A specific example of a data analysis service is classifying different types of water quality and predicting (forecasting) how water quality will change over time. Finally, with the above cloud 4IR platform deployed, data from the WWTP is displayed on a web page, allowing remote users to perform and control water quality analysis. Figure 9.2 is a detailed view of the EDAR 4.0, 4IR system architecture. The diagram also describes the software tools used for cloud 4IR components. For this work, the Flask library API with Python tools was used. The PostgreSQL database was used for storage. The Rapidminer program was used for the analysis. Finally, the Bokeh library was used for the rendering part.

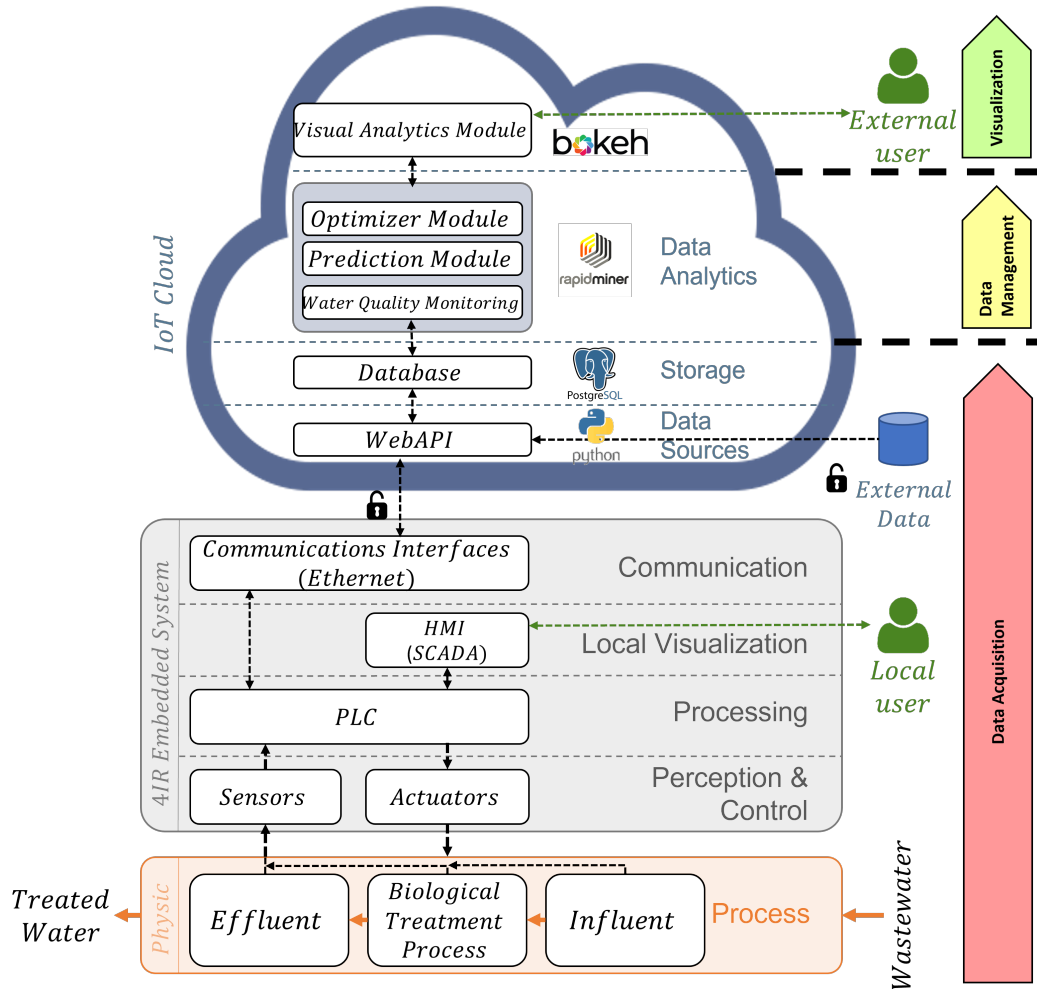


Figure 9.2: Proposed EDAR 4.0 architecture.

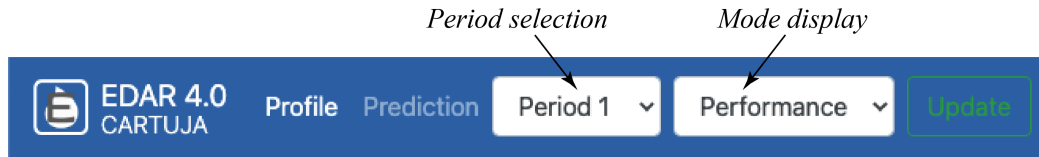
## 9.3 Innovation

As a result of following the EDAR 4.0 architecture, which was based on the AvRuskin methodology, the software tool “EDAR 4.0” was created. In the following, its modules and a discussion of its validation with the end-user are detailed.

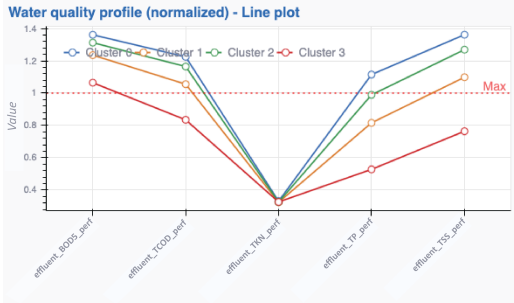
### 9.3.1 Water quality monitoring

The dataset generated by the WWTP “La Cartuja” SCADA system was subjected to a series of steps to preprocess it and leave it ready for the Data Cleaning process. After cleaning the data, the Principal Component Analysis (PCA) method was applied to extract the two main components that define the dataset. In addition, the clustering process was performed using the K-means algorithm with  $k=4$ . Each group identified by the algorithm belongs to a water quality cluster. The platform allowed parameterising if the water quality monitoring was displayed on water treatment per-

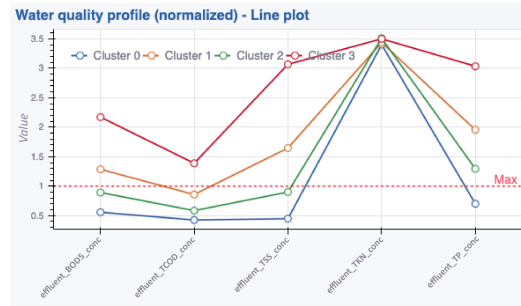
formance or absolute values in the front end. Another parameter that users could specify from the platform was the duration of the treatment plant. This was done because the processing plant “La Cartuja” had improved its equipment. Therefore, it was essential to track and separate these two periods. Water quality profiles (or clusters) were constructed using line profiles and spider plots. Figure 9.3 shows the monitoring module of the EDAR 4.0 platform. This graph shows that the worst water quality is in the blue cluster (cluster 0), and the best quality is in the red cluster (cluster 3). Also, it should be noted that the WWTP “La Cartuja” needs to handle NTK chemical variables better.



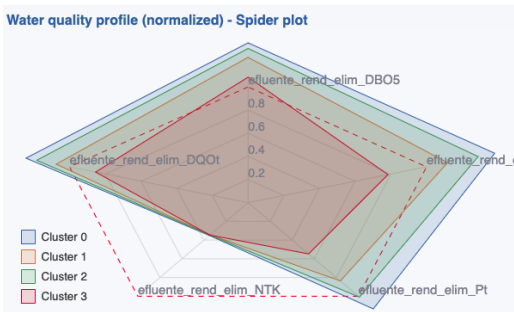
(a) Monitoring configuration parameters



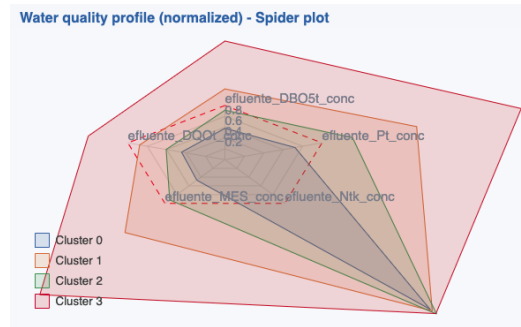
(b) Water Quality Line Chart (Performance)



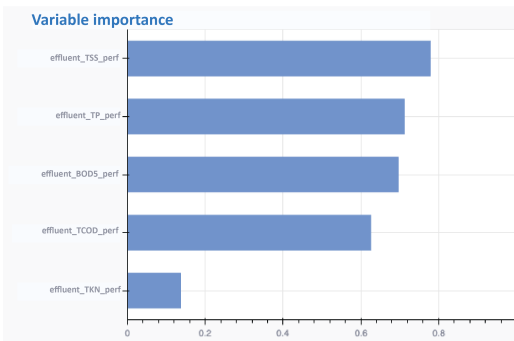
(c) Water Quality Line Chart (Absolute)



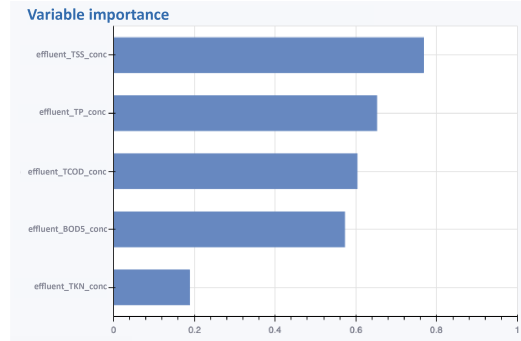
(d) Water Quality Spider Chart (Performance)



(e) Water Quality Spider Chart (Absolute)



(f) Water Quality Variable Importance (Performance)

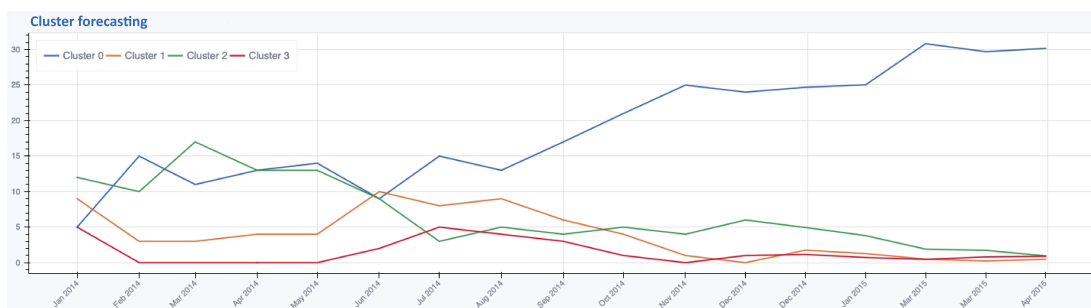


(g) Water Quality Variable Importance (Absolute)

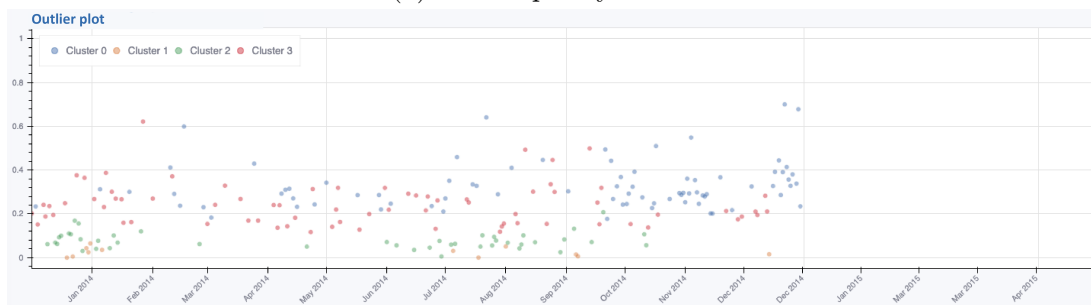
Figure 9.3: Visual Analytics Water Quality Monitoring Platform.

### 9.3.2 Water quality prediction

The water quality prediction tool predicted the number of months the wastewater treatment plant would have each water quality cluster. The method used was a Holt-Winters time series forecasting implemented in the back end. The front end plotted two graphs: i) Time Series Cluster Prediction Plot and ii) an Outlier Probability Plot (see Figure 9.4). The vertical dotted line separates the data set from the forecast data. Wastewater treatment plant operators should ideally be in the red cluster (cluster 3) with the highest predicted values for the best water quality on this graph and in the blue cluster (cluster 0) with the lowest predicted values for the worst water quality. Note that the forecast graph shows values above 30. This is because the forecasts do not consider that there are 30 days in a month.



(a) Water quality forecast



(b) Water quality forecast outlier probability plot

Figure 9.4: Visual Analytics Water Quality Prediction Platform.

### 9.3.3 WWTP Model Creation & Simulation

At this stage, a database model of the energy or chemical variables of the wastewater treatment process was created. The default model generated by the platform was water quality. For example, energy consumption (kilowatts per day) can be modelled as a function of all other process variables. On the back end, the implemented machine learning system was able to determine the most relevant variables for modelling based on the correlation matrix. The method used to create the model was decision trees. Once the model was created, it was possible to interact with the associated platform variables. Once the values were selected, predicting the modelled variable's range of

values would be with those with which the model was simulated. This process is shown in Figure 9.5 using as an example the electricity consumption modelling, where a set of values is given for the relevant values and after simulating the platform predicts that the WWTP will be at a range1 ( $-\infty$  to 59816 kW) of energy consumption.

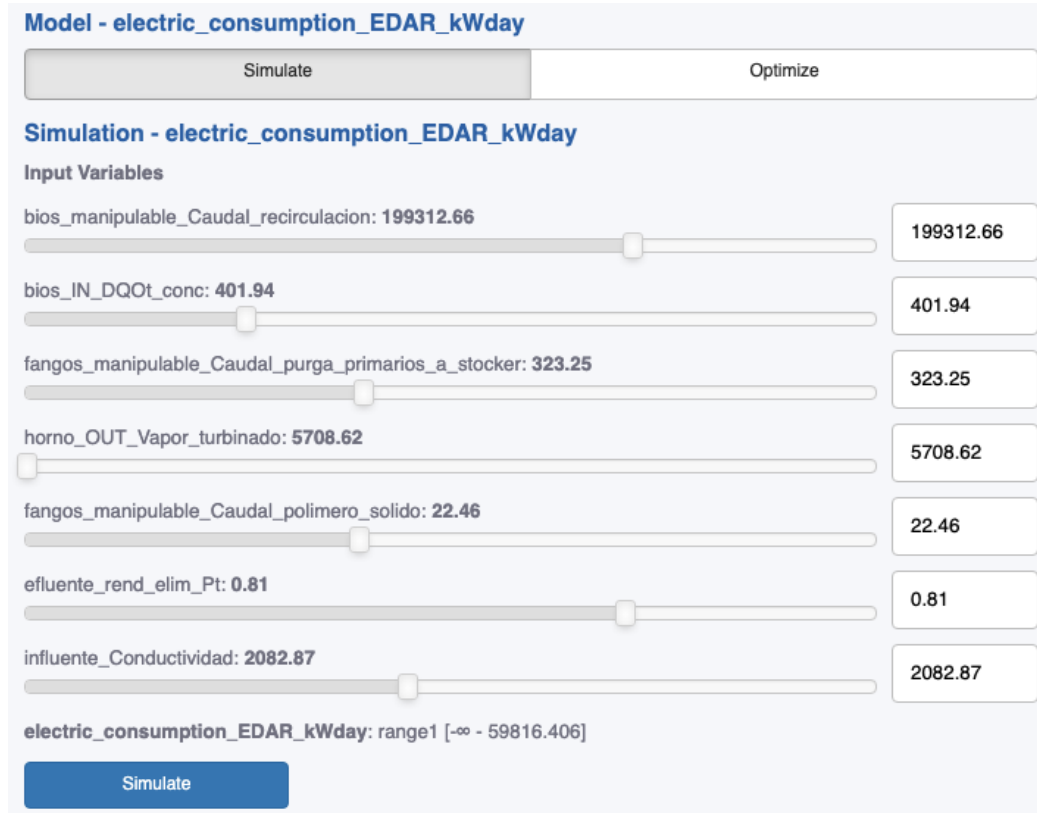


Figure 9.5: Energy consumption model simulation.

The confusion matrix allows the visualisation of the model performance, which is displayed in Figure 9.6. This shows how many of the values predicted by the model were correct according to the label. In addition, the developed platform allows the operator to check the importance of variables in the created model (see Figure 9.7). Finally, the dashboard displays the decision tree generated for the given variable, as shown in Figure 9.8.

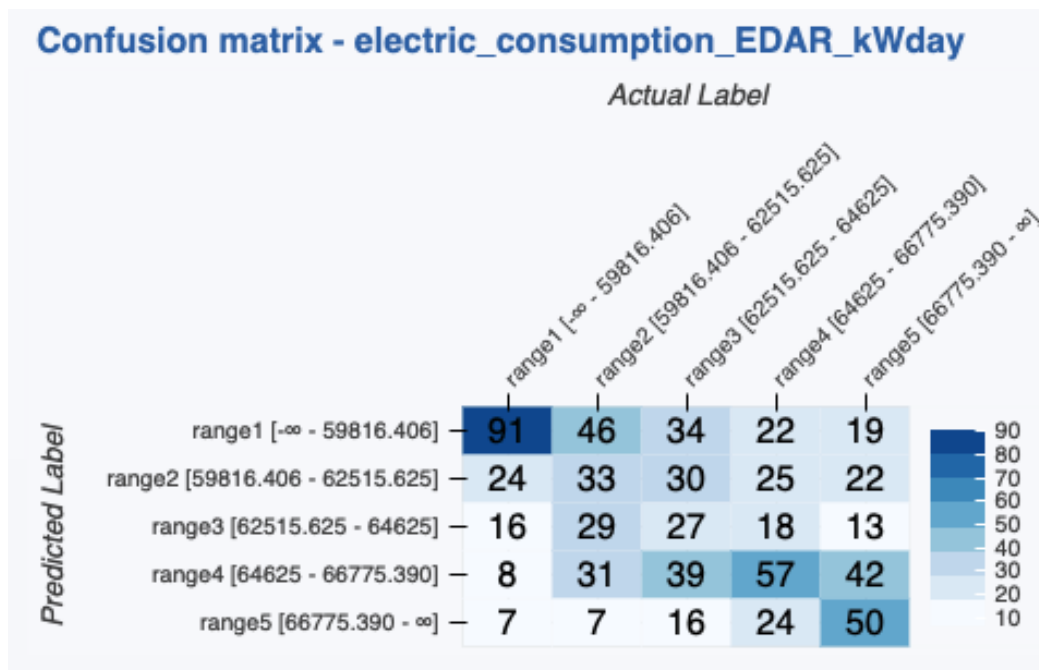


Figure 9.6: Confusion matrix for electric model.

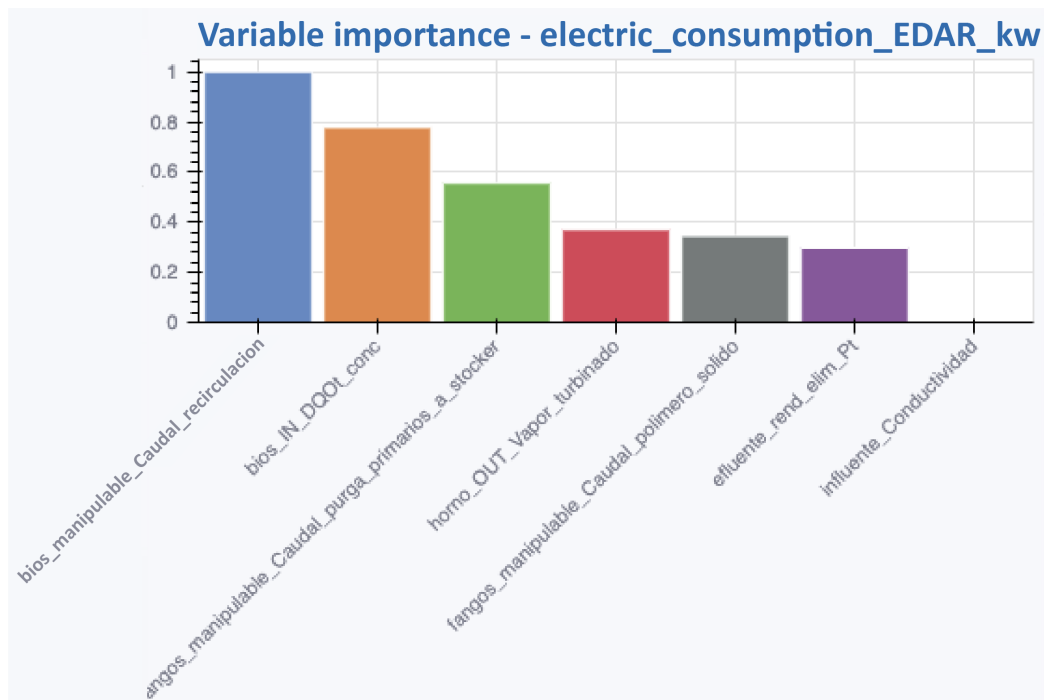


Figure 9.7: Variable influence for electric model.

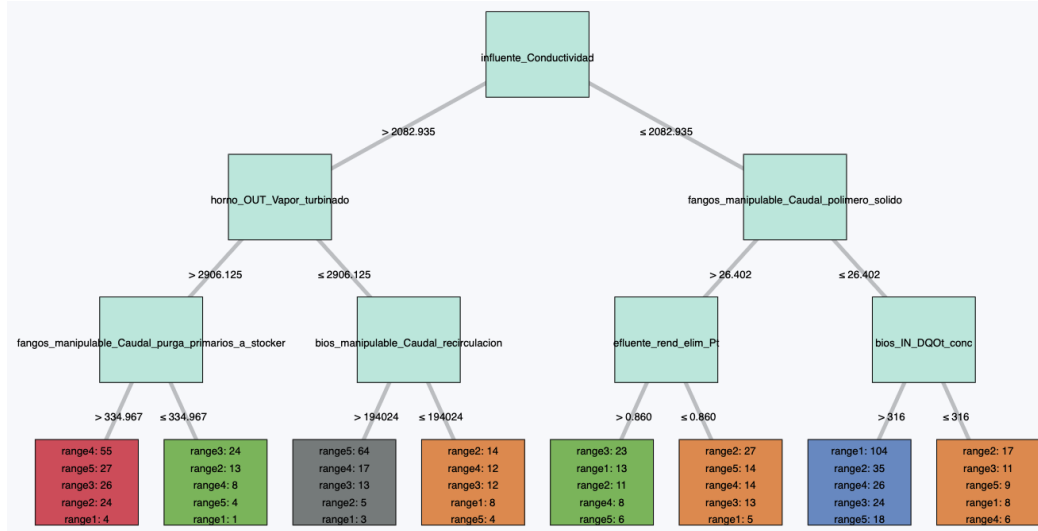


Figure 9.8: Decision tree for electric model.

### 9.3.4 WWTP Model Optimisation

This component of the platform works oppositely to the simulator, where a target interval (range) is set for the variable being modelled, and restrictions are placed on the variables that influence it. Once this has been done, optimal values can be obtained for each influential variable to guarantee the modelled variable’s target with the given restrictions. As an example, it could be found in Figure 9.9 which are the values of the chemical concentrations to be used to obtain the lowest possible range of energy consumption for the WWTP.



**Optimization - electric\_consumption\_EDAR\_kWday**

Objective: Maximize Target: range1 [-∞ - 59816.406]

**Constraints**

bios_manipulable_Caudal_recirculacion: range1 [-∞ - 190044]	Condition 1 -	Value 1 range2 ['
bios_IN_DQOt_conc: range3 [357 - 421]	Condition 1 -	Value 1 range3 [;
fangos_manipulable_Caudal_purga_primarios_a_stocker: range1 [-∞ - 257.734]	Condition 1 -	Value 1 range3 [;
horno_OUT_Vapor_turbinado: range3 [2699.021 - 3276.177]	Condition 1 -	Value 1 range1 [-
fangos_manipulable_Caudal_polimero_solido: range3 [17.926 - 27.214]	Condition 1 -	Value 1 range1 [-
efluente_rend_elim_Pt: range5 [0.903 - ∞]	Condition 1 -	Value 1 range2 [(
influyente_Conductividad: range1 [-∞ - 1659.500]	Condition 1 -	Value 1 range2 ['

Prediction: range1 [-∞ - 59816.406], Confidence: 94.602 %

**Optimize**

Figure 9.9: Energy consumption model optimisation.

### 9.3.5 User's validation

End users have confirmed the operational improvements made by the developed tools. This improvement includes the following aspects of the current tools:

- **Observability:** it allows monitoring the state of water quality through a visualisation based on clustering.
- **Predictability:** the operator can forecast how his WWTP will go in the future.
- **Risk-free evaluation:** operators can validate how their system will perform if specific parameters change through simulation and optimisation. This represents an essential advantage because, currently, they were required to test their actual WWTP, which could lead to damage if their operating variables were manipulated.
- **Interpretability:** The decision trees and variable importance graphs helped the operator better understand his WWTP behaviour.

Users recognise that these benefits can be obtained without needing well-trained and qualified personnel. Although this aspect can be interpreted as limiting, it is ultimately viewed as positive by users as continuing education and training are part of employees' rights and the company's obligations. So this is seen as a possibility, not a limitation.

Finally, in addition to this qualitative validation, it has been possible to perform a quantitative validation of the quality of the tools models. Specifically, as it can be seen in the confusion matrix on Figure 9.10.

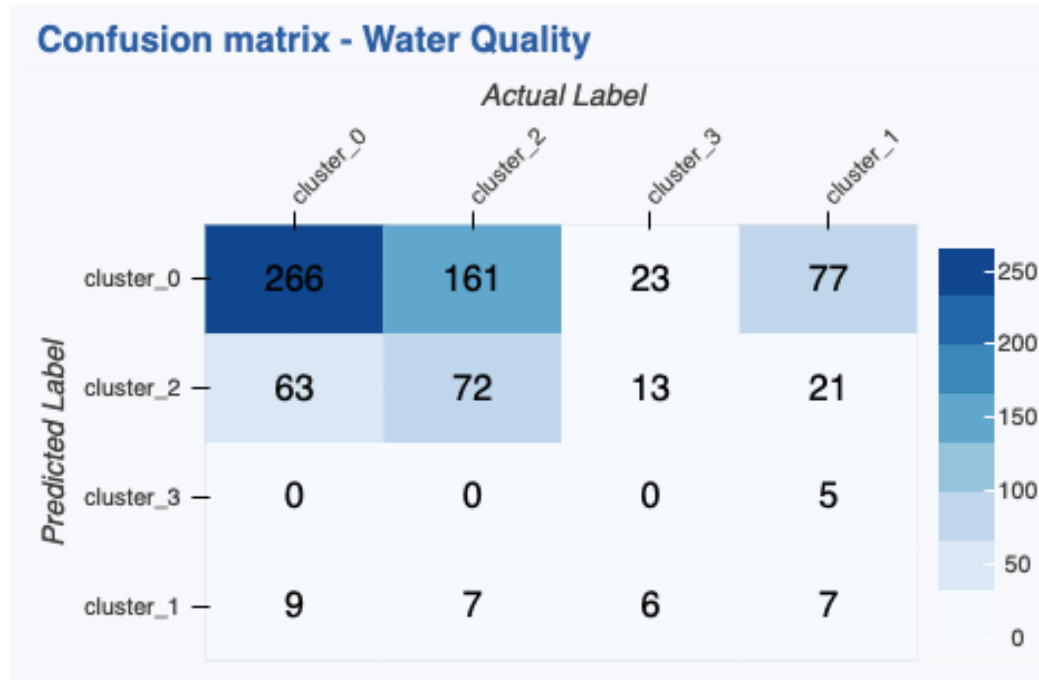


Figure 9.10: Confusion matrix for water quality model.

Furthermore, the predictor importance plot (see Figure 9.11) shows the variables that have the most significant impact on the operation of the treatment plant, according to the model built and validated by end users. It was found that these variables do have a direct effect on wastewater quality, which is the best evidence to confirm the results.

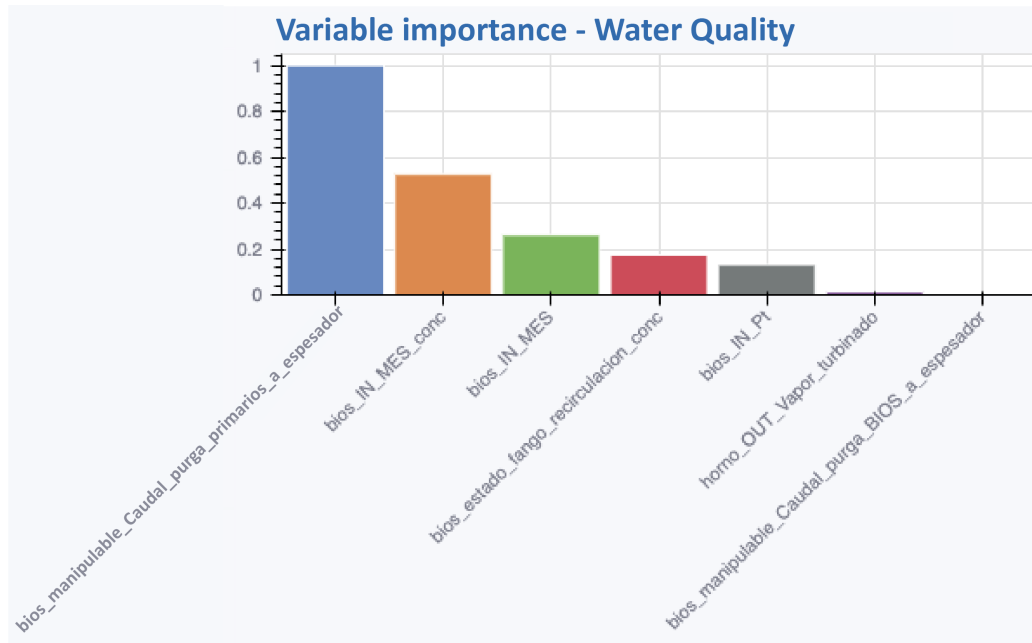


Figure 9.11: Variable importance for water quality model.

## 9.4 Conclusions and future work

This chapter described a visual analytics-based wastewater treatment plant platform called EDAR 4.0. Intuitive visualisations have great potential to support decision-making in the operation and management of WWTPs. The proposed tool allows users to discover relationships between data through simple data validation. The developed tool allows wastewater plant operators to perform modelling and optimisation without compromising real-world field testing. This tool has been endorsed by WWTP experts and has shown to be an additional source of information for WWTP management. For future work, it is suggested to first scale the tool for a multi-factory approach. In addition, this tool can be used for the dynamic monitoring of ammonium, which is a novelty for WWTPs. It is also proposed to carry out an in-depth study concerning usability.

---

## Conclusions and future work

---

In this chapter the final conclusions of the present research are presented and the proposed future work.

### 10.1 Conclusions

As presented in the previous chapters, the Fourth Industrial Revolution (4IR) has brought disruptive technologies to enhance existing industrial systems or create from scratch systems with these new capabilities. One of the problems faced by industries is to be able to quickly detect any anomalies in their systems in order to make the respective corrections or decisions. For this reason, this thesis aims to provide a framework to guide data engineers in designing and constructing 4IR-enabled systems for anomaly detection. Detecting anomalies in real-time is a challenging task. One of the problems encountered in this process is that information must be processed carefully and quickly to provide a reliable and early response to an anomaly. Another common problem is the accuracy of prediction systems, where a Machine Learning algorithm may perform better at predicting anomalies than others. There may be missing data in the acquisition of data used to detect these anomalies, so it is crucial to have strategies to tackle this issue. Finally, presenting the prediction information of these Artificial Intelligence (AI) driven anomaly detection systems to the end user so that in a clear and understandable way is also a challenge. This research sought to answer the research questions proposed in Chapter 1, background.

The present research proposed a new hardware and software architecture for AI-driven Industry 4.0 systems inspired by state-of-the-art architectures and three real industrial use cases. This architecture includes a physical layer, an embedded system layer and an Internet of Things (IoT) cloud layer, providing the users with a clear view and practical guidelines for including the components required to implement a 4IR system.

Then, the design and construction process of a test bench cyber-physical data acquisition system that integrates Remote Sensing (RS) and Wireless Sensor Networks (WSN) for detecting pests in coffee crops was presented. Within this research work, it was possible to create a 3-month dataset containing different sources of information, particularly images from RGB, multispectral cameras and data from precision agriculture sensors. The data acquisition process is the first part of a Machine Learning algorithm's pipeline for detecting possible anomalies. This result explains how to reliably obtain information to subsequently create predictive Machine Learning models, which in this case made it possible to detect the level of Coffee Leaf Rust disease in coffee crops.

Next, a supervised anomaly detection system for the detection of Coffee Leaf Rust was introduced. The novelty of this system is that it integrates multiple sources of information, specifically Remote Sensing, Wireless Sensor Networks and Deep Learning for the creation of an ensembled Machine Learning meta-model. With this result, it was possible to predict the severity of the Coffee Leaf Rust disease, which is considered an anomaly in the coffee plant. In this particular case, it is shown that combining models does not necessarily improve the overall anomaly prediction performance since using the multispectral cameras or the RGB camera on their own gave an F1-score of greater than 0.9, whereas using the combined model gave an F1-score of 0.775. This can be explained by the fact that the data source from the Precision Agriculture sensors was very noisy or had a significant amount of missing information, thus impairing the system's overall response. Nevertheless, it was proved that the performance of an AI-driven anomaly detection system is comparable and reliable enough to replace human visual inspection.

It is common to find that industrial systems do not have a ground truth to distinguish when the system is faulty or in a normal state. In these cases, semi-supervised techniques can help generating preliminary labels to train a subsequent Machine Learning model. In this research, a Hybrid Machine-Learning Ensemble for Anomaly Detection for a real-time Industry 4.0 system was developed and created. This system was inspired by the process of an industrial machine, thus having a Manufacturing Stage and an Operation Stage. This system was tested on three different machines, obtaining F1-scores on higher than 0.9 on average. This ensembling consisted of three state-of-the-art algorithms for anomaly detection: Autoencoder, One Class Support Vector Machine and Local Outlier Factor. The ensemble model improves over the individual models for detecting anomalies in real-time. It is concluded that in some industrial cases, it is beneficial to have multiple decision sources for anomaly detection since a Machine Learning algorithm can better detect specific patterns in some instances and having different algorithms allows to cover more cases where the anomaly is present.

Anomalies can be challenging to detect with the naked eye using the time domain. This is why using frequency domain transformations can provide additional information about the behaviour of an industrial system. This research performed a frequency domain analysis for cracks in compacted hygroscopic material. In this case, the spectrogram method and a vibratory test bench were used to identify the exact time point where the compacted samples failed (had a crack). With this, it

was possible to simulate the transport conditions to make decisions on the design of the compacted material to be transported, such as its compression pressure or the design of the truck's dampers that must be used for transporting these materials. It is noted that the anomalies in the frequency domain were identified when new frequency components appeared in addition to the primary vibration frequency.

Finally, the data from the previously created models must be visualised clearly and efficiently in the machine learning process. This research proposed a Visual Analytics platform for water quality monitoring in a Wastewater Treatment Plant (WWTP). In this case, this platform allowed the creation of Machine Learning models to identify different water qualities and to simulate wastewater treatment processes. The developed tools allowed the creation, simulation and optimisation of Machine Learning models for any variable of the WWTP process. Domain experts validated this platform. The end-user was thus able to test different parameters of the WWTP process to guarantee a specific water quality and/or to optimise the energy management of the WWTP.

## 10.2 Future work

Each of the contributions presented in this thesis introduces challenges for future work. In the case of 4IR architectures, the challenge is to create a tool to guide the end-user in creating a specific architecture for their industrial context. Each component of this architecture should be based on the user's requirements. Other industrial use cases could be analysed to extend the proposed generic architecture.

In the data acquisition process, extending the current data collection system's design with costing studies, scalability analysis, and energy consumption study is relevant to convert from a laboratory test bed into a full-scale data acquisition system.

Concerning the case study of supervised ensembling methods for anomaly detection, it is proposed to test other Machine Learning algorithms to improve the ensembling results, as future work. In this case, the precision agriculture sensors' data source was the worst performer, so this could be further improved with other Machine Learning algorithms. It is also proposed to test missing data imputation techniques to obtain a more complete data set.

In the research work carried out in chapter 7, semi-supervised ensembling techniques for anomaly detection, it is proposed to explore other algorithms, such as Isolation Forest and the Elliptic Envelop, to improve the overall performance of the models, as future work. Furthermore, it is proposed to study techniques for the adaptability of the Machine Learning model as it wears out and loses validity over time. For this, for example, techniques for retraining Machine Learning algorithm models can be further explored. In addition, it is noted that the data from the industrial machines used in this case study had a significant amount of null or missing data. Thus, exploring further data imputation techniques is also proposed.

For the method presented in this thesis for detecting cracks (anomalies) in compacted hygroscopic particulate materials, it is proposed to complement the implemented algorithm with artificial vision techniques that allow recognising the exact

point (location) where the fracture occurred, as future work, in order to make the proposed algorithm more robust.

Finally, the Visual Analytics techniques presented for single Wastewater Treatment Plants can be completed by addressing a multiple plant approach. Additionally, the proposed visualisation tools can be extended with other explanatory techniques of Machine Learning models, e.g., analysing the relationships between process and/or operational variables of the WWTP, such that clearer, more understandable and richer information is provided to the users.

---

## Bibliography

---

- [1] Shikha Agrawal and Jitendra Agrawal. Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, 60:708–713, 2015. Knowledge-Based and Intelligent Information & Engineering Systems 19th Annual Conference, KES-2015, Singapore, September 2015 Proceedings.
- [2] Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, 2017.
- [3] Chuadhry Mujeeb Ahmed, Gauthama Raman M R, and Aditya P. Mathur. Challenges in machine learning based approaches for real-time anomaly detection in industrial control systems. In *Proceedings of the 6th ACM on Cyber-Physical System Security Workshop*, CPSS '20, page 23–29, New York, NY, USA, 2020. Association for Computing Machinery.
- [4] Rama Al-Attar, Mouhammd Alkasassbeh, and Mu'Awya Al-Dala'Ien. A survey: Soft computing for anomaly detection to mitigate iot abuse. In *2022 International Conference on Engineering & MIS (ICEMIS)*, pages 1–6, 2022.
- [5] Tsatsral Amarbayasgalan, Bilguun Jargalsaikhan, and Keun Ho Ryu. Unsupervised novelty detection using deep autoencoders with density based clustering. *Applied Sciences (Switzerland)*, 2018.
- [6] Sang Ha An, Gyunyoung Heo, and Soon Heung Chang. Detection of process anomalies using an improved statistical learning framework. *Expert Systems with Applications*, 2011.
- [7] Mauricio Aramburo-Londoño, Santiago Pérez-Cardona, Manuela Calle-Escobar, Alejandro Velásquez-López, and Ricardo Mejía-Gutiérrez. Impact analysis of compressed hygroscopic particulate material. *International Journal of Mechanical and Mechatronics Engineering*, 2016.



- [8] Gillian A AvRuskin, Geoffrey M Jacquez, Jaymie R Meliker, Melissa J Slotnick, Andrew M Kaufmann, and Jerome O Nriagu. Visualization and exploratory analysis of epidemiologic data using a novel space time information system. *International Journal of Health Geographics*, 3(1):26, 2004.
- [9] Francesco Aymerich, Wieslaw J Staszewski, and Tadeusz Uhl. Effect of boundary conditions on nonlinear acoustics used for impact damage detection in composite structures. In Tribikram Kundu, editor, *Health Monitoring of Structural and Biological Systems 2010*, volume 7650, pages 934–943. International Society for Optics and Photonics, SPIE, 2010.
- [10] Saeed Azfar, Adnan Nadeem, and Abdul Basit Shaikh. Pest Detection and Control Techniques Using Wireless Sensor Network: A Review. *Journal of entomology and zoology studies*, 3:92–99, 2015.
- [11] Behrad Bagheri, Shanhu Yang, Hung An Kao, and Jay Lee. Cyber-physical systems architecture for self-aware machines in industry 4.0 environment. In *IFAC-PapersOnLine*, 2015.
- [12] V. Barnett and T. Lewis. *Outliers in statistical data*. John Wiley & Sons Ltd., 2nd edition edition, 1978.
- [13] Massimo Blonda, Angelantonio Calabrese, Angelo Cardellicchio, Barbara Casale, Giuseppe Dentamaro, Vincenzo Di Lecce, Antonietta Dimucci, Cataldo Guaragnella, Diego Matrino, Dian Palagachev, Domenico Petruzzelli, Tiziano Politi, Maria Rizzi, Vincenzo Sarcina, and Vito Felice Uricchio. Innovative Methodology for Detecting of Possible Harmful Compounds for Wastewater Treatment the MAUI Project. In *2018 Workshop on Metrology for Industry 4.0 and IoT, MetroInd 4.0 and IoT 2018 - Proceedings*, 2018.
- [14] Julian Andrés Bolaños, Liseth Campo, and Juan Carlos Corrales. Characterization in the Visible and Infrared Spectrum of Agricultural Crops from a Multi-rotor Air Vehicle. In *International Conference of ICT for Adapting Agriculture to Climate Change*, pages 29–43. Springer, 2017.
- [15] Azzedine Boukerche, Lining Zheng, and Omar Alfandi. Outlier Detection: Methods, Models, and Classification. *ACM Computing Surveys*, 2020.
- [16] Markus M. Breunig, Hans Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 2000.
- [17] Gabriela Calvario, Basilio Sierra, Teresa E. Alarcón, Carmen Hernandez, and Oscar Dalmau. A multi-disciplinary approach to remote sensing through low-cost UAVs. *Sensors (Switzerland)*, 2017.
- [18] A. Camargo and J. S. Smith. An image-processing based algorithm to automatically identify plant disease visual symptoms. *Biosystems Engineering*, 2009.

- [19] A. T. Catherall and D. P. Williams. High resolution spectrograms using a component optimized short-term fractional Fourier transform. *Signal Processing*, 90(5):1591–1596, 2010.
- [20] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), July 2009.
- [21] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection for discrete sequences: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 24:1–1, 05 2012.
- [22] Varun Chandola, Varun Mithal, and Vipin Kumar. Comparative evaluation of anomaly detection techniques for sequence data. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2008.
- [23] Yung-Chun Chang, Chih-Hao Ku, and Chien-Hung Chen. Using deep learning and visual analytics to explore hotel reviews and responses. *Tourism Management*, 80:104129, 2020.
- [24] D D Chaudhary, S P Nayse, and L M Waghmare. Application of wireless sensor networks for greenhouse parameter control in precision agriculture. *International Journal of Wireless & Mobile Networks (IJWMN)*, 3(1):140–149, 2011.
- [25] Abel Chemura, Onesimo Mutanga, and Timothy Dube. Remote sensing leaf water stress in coffee (*Coffea arabica*) using secondary effects of water absorption and random forests. *Physics and Chemistry of the Earth, Parts A/B/C*, 100:317–324, 2017.
- [26] Cailian Chen, Jing Yan, Ning Lu, Yiyin Wang, Xian Yang, and Xinping Guan. Ubiquitous monitoring for industrial cyber-physical systems over relay-assisted wireless sensor networks. *IEEE Transactions on Emerging Topics in Computing*, 3(3):352–362, 2015.
- [27] M. Chugani, A. Samant, and M. Cerna. *LabVIEW Signal Processing*. Prentice Hall, New York, USA, 1st edition, 1998.
- [28] Andrew A. Cook, Göksel Mısırlı, and Zhong Fan. Anomaly detection for iot time-series data: A survey. *IEEE Internet of Things Journal*, 7(7):6481–6494, 2020.
- [29] Hong Ning Dai, Hao Wang, Guangquan Xu, Jiafu Wan, and Muhammad Imran. Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies. *Enterprise Information Systems*, 2019.
- [30] Robert P.W. Duin David M.J. Tax. Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research*, 2002.

- [31] Elias De Melo Virginio Filho and Carlos Astorga. *Prevención y control de la roya del café: Manual de buenas prácticas para técnicos y facilitadores*. CATIE, Turrialba, C.R, 1 edition, 2015.
- [32] Nan Ding, Hao Xuan Ma, Huanbo Gao, Yan Hua Ma, and Guo Zhen Tan. Real-time anomaly detection based on long short-Term memory and Gaussian Mixture Model. *Computers and Electrical Engineering*, 2019.
- [33] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 edition, November 2000.
- [34] P. Duhamel and M. Vetterli. Fast fourier transforms: A tutorial review and a state of the art. *Signal Processing*, 19(4):259–299, 1990.
- [35] Issam El Naqa and Martin J. Murphy. *What Is Machine Learning?*, pages 3–11. Springer International Publishing, Cham, 2015.
- [36] Muhammad Fahim and Alberto Sillitti. Anomaly Detection, Analysis and Prediction Techniques in IoT Environment: A Systematic Literature Review, 2019.
- [37] B. G. Ferguson. Time-frequency signal analysis of hydrophone data. *Oceanic Engineering, IEEE Journal of*, 21(4):537–544, 1996.
- [38] Susana Ferreira, Basilio Sierra, Itziar Irigoien, and Eneko Gorritxategi. A bayesian network for burr detection in the drilling process. *J. Intell. Manuf.*, 23(5):1463–1475, 2012.
- [39] Xavier Flores-Alsina, Christian Kazadi Mbamba, Kimberly Solon, Darko Vrecko, Stephan Tait, Damien J. Batstone, Ulf Jeppsson, and Krist V. Gernaey. A plant-wide aqueous phase chemistry module describing ph variations and ion speciation/pairing in wastewater treatment process models. *Water Research*, 85:255–265, 2015.
- [40] Alvaro Fuentes, Sook Yoon, Sang Cheol Kim, and Dong Sun Park. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors (Switzerland)*, 2017.
- [41] Achim Gahr, Peter Wazinski, and Nils Andreas. Water Management 4.0 in the Bitterfeld-Wolfen Chemical Park. *Chemie-Ingenieur-Technik*, 2019.
- [42] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Hamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, 46, 04 2014.
- [43] Raghu K. Ganti, Fan Ye, and Hui Lei. Mobile crowdsensing: Current state and future challenges. *IEEE Communications Magazine*, 2011.
- [44] Jiechao Gao, Haoyu Wang, and Haiying Shen. Task failure prediction in cloud data centers using deep learning. *IEEE Transactions on Services Computing*, PP:1–1, 05 2020.

- [45] Nicolás García-Pedrajas, César Hervás-Martínez, and Domingo Ortiz-Boyer. Cooperative coevolution of artificial neural network ensembles for pattern classification. *IEEE transactions on evolutionary computation*, 9(3):271–302, 2005.
- [46] Gilbert-Rainer Gillich and Zeno-Iosif Praisach. Modal identification and damage detection in beam-like structures using the power spectrum and time–frequency analysis. *Signal Processing*, 96, Part A:29–44, 2014.
- [47] Omid Givehchi, Henning Trsek, and Juergen Jasperneite. Cloud computing for industrial automation systems — a comprehensive overview. In *2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA)*, pages 1–4. IEEE, 2013.
- [48] Pradeep K Goel, Shiv O Prasher, Jacques-André Landry, Ramanbhai M Patel, R B Bonnell, Alain A Viau, and J R Miller. Potential of airborne hyperspectral remote sensing to detect nitrogen deficiency and weed infestation in corn. *Computers and electronics in agriculture*, 38(2):99–124, 2003.
- [49] H. M. Gomes, D. dos Santos Gaspareto, F. de Souza Ferreira, and C. A. K. Thomas. A Simple Closed-Loop Active Control of Electrodynamic Shakers by Acceleration Power Spectral Density for Environmental Vibration Tests. *Experimental Mechanics*, 48(5):683–692, Oct 2008.
- [50] GSMA Association. Understanding the Internet of Things (IoT). *Gsma Connected Living*, page 15, 2014.
- [51] Fernando Haddad, Luiz A Maffia, Eduardo SG Mizubuti, and Hudson Teixeira. Biological control of coffee rust by antagonistic bacteria under field conditions in brazil. *Biological Control*, 49(2):114–119, 2009.
- [52] Esmael Hamuda, Martin Glavin, and Edward Jones. A survey of image processing techniques for plant extraction and segmentation in the field. *Computers and Electronics in Agriculture*, 125:184–199, 2016.
- [53] D. M. Hawkins. *Identification of outliers*. Monographs on applied probability and statistics. Chapman and Hall, London [u.a.], 1980.
- [54] Geoffrey Hinton and Terrence J. Sejnowski. *Unsupervised Learning: Foundations of Neural Computation*. The MIT Press, 05 1999.
- [55] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [56] M Hubert and E Vandervieren. An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, 52(12):5186–5201, 2008.
- [57] Geoff Hulten, Laurie Spencer, and Pedro Domingos. Mining time-changing data streams. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001.

- [58] Hanan Hussain, PS Tamizharasan, and CS Rahul. Design possibilities and challenges of dnn models: a review on the perspective of end devices. *Artificial Intelligence Review*, pages 1–59, 2022.
- [59] Industrial Internet Consortium. The Industrial Internet Reference Architecture v1.9, 2019.
- [60] Intel Corporation. Connecting legacy devices to the Internet of Things, 2014.
- [61] U. Jeppsson, C. Rosen, J. Alex, J. Copp, K. V. Gernaey, M.-N. Pons, and P. A. Vanrolleghem. Towards a benchmark simulation model for plant-wide control strategy performance evaluation of wwtps. *Water Science and Technology*, 53(1):287–295, Jan 2006.
- [62] Tammy Jiang, Jaimie L Gradus, and Anthony J Rosellini. Supervised machine learning: a brief primer. *Behavior Therapy*, 51(5):675–687, 2020.
- [63] Kim Jongrack, You Kwangtae, Piao Wenhua, and Kim Yejin. Modified newton-raphson method to minimize calculation time for wastewater treatment plant simulation. *J. Korean Soc. Hazard Mitig*, 18(5):319–326, 2018.
- [64] David Jonker, Richard Brath, and Scott Langevin. Industry-driven visual analytics for understanding financial timeseries models. In *2019 23rd International Conference Information Visualisation (IV)*, pages 210–215. IEEE, 2019.
- [65] A. Jović, K. Brkić, and N. Bogunović. A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1200–1205, 2015.
- [66] JRC of the European Commission. Precision Agriculture: an Opportunity for Eu Farmers- Potential Support With the Cap 2014 - 2020. *European Union*, page 56, 2014.
- [67] Hairulliza Judi, Ruzzakiah Jenal, and Devendran Genasan. *Quality Control Implementation in Manufacturing Companies: Motivating Factors and Challenges*, chapter 25. IntechOpen, 04 2011.
- [68] M. J. Jweeg, E. Q. Hussein, and K. I. Mohammed. Effects of cracks on the frequency response of a simply supported pipe conveying fluid. *International Journal of Mechanical and Mechatronics Engineering*, 2017.
- [69] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. *Visual Analytics: Definition, Process, and Challenges*, pages 154–175. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
- [70] Siddhartha Kumar Khaitan and James D. McCalley. Design techniques and applications of cyberphysical systems: A survey. *IEEE Systems Journal*, 2015.

- [71] Zafran Khan, Naima Iltaf, Hammad Afzal, and Haider Abbas. Enriching non-negative matrix factorization with contextual embeddings for recommender systems. *Neurocomputing*, 380:246–258, 2020.
- [72] M. Kim, Y. Kim, H. Kim, W. Piao, and C. Kim. Operator decision support system for integrated wastewater management including wastewater treatment plants and receiving water bodies. *Environ Sci Pollut Res Int*, 23(11):10785–10798, Jun 2016.
- [73] Josef Kittler, Mohamad Hatf, Robert PW Duin, and Jiri Matas. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 1998.
- [74] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.
- [75] Nikolay Laptev, Saeed Amizadeh, and Ian Flint. Generic and scalable framework for automated time-series anomaly detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.
- [76] Edward A. Lee. Cyber physical systems: Design challenges. In *Proceedings - 11th IEEE Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing, ISORC 2008*, 2008.
- [77] Jay Lee, Behrad Bagheri, and Hung An Kao. A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters*, 2015.
- [78] Min Hwa Lee, Jin Hyo Joseph Yun, Andreas Pyka, Dong Kyu Won, Fumio Kodama, Giovanni Schiuma, Hang Sik Park, Jeonghwan Jeon, Kyung Bae Park, Kwang Ho Jung, Min Ren Yan, Sam Youl Lee, and Xiaofei Zhao. How to respond to the Fourth Industrial Revolution, or the second information technology revolution? Dynamic new combinations between technology, market, and society through open innovation. *Journal of Open Innovation: Technology, Market, and Complexity*, 2018.
- [79] F. Leonard. Phase spectrogram and frequency spectrogram as new diagnostic tools. *Mechanical Systems and Signal Processing*, 21(1):125–137, 2007.
- [80] F. Leonard, J. Lanteigne, S. Lalonde, and Y. Turcotte. Free-vibration behaviour of a cracked cantilever beam and crack detection. *Mechanical Systems and Signal Processing*, 15(3):529–548, 2001.
- [81] Marianna Lezzi, Mariangela Lazoi, and Angelo Corallo. Cybersecurity for industry 4.0 in the current literature: A reference framework. *Computers in Industry*, 103:97–110, 2018.

- [82] Ping Li, Jianping Li, and Gongcheng Wang. Application of convolutional neural network in natural language processing. In *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 120–122. IEEE, 2018.
- [83] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations, 2020.
- [84] Fei Tony Liu, Kai Ming Ting, and Zhi Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data*, 2012.
- [85] Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, 1(1):48–56, 2017.
- [86] Ioannis E Livieris, Emmanuel Pintelas, and Panagiotis Pintelas. A cnn-lstm model for gold price time-series forecasting. *Neural computing and applications*, 32(23):17351–17360, 2020.
- [87] Brad Lobitz, Louisa Beck, Anwar Huq, Byron Wood, George Fuchs, A S G Faruque, and Rita Colwell. Climate and infectious disease: use of remote sensing for detection of *Vibrio cholerae* by indirect measurement. *Proceedings of the National Academy of Sciences*, 97(4):1438–1443, 2000.
- [88] J. F. Macgregor. On-line statistical process control. *Chemical Engineering Progress*, 84:21–31, 1988.
- [89] M Maiza, J Odriozola, A Gil, G Naveran, R Basagoiti, I Lecuona, U Zurutuza, G Urchehi, and A Mañas. Visual analytics for supporting the management of wwtps. In *Proceedings of the Young Water Professionals (YWP) conference*, 2017.
- [90] Federico Martinelli, Riccardo Scalenghe, Salvatore Davino, Stefano Panno, Giuseppe Scuderi, Paolo Ruisi, Paolo Villa, Daniela Stroppiana, Mirco Boschetti, Luiz R. Goulart, Cristina E. Davis, and Abhaya M. Dandekar. Advanced methods of plant disease detection. a review. *Agronomy for Sustainable Development*, 35(1):1–25, Jan 2015.
- [91] Evgenii S Matrosov, Ivana Huskova, Joseph R Kasprzyk, Julien J Harou, Chris Lambert, and Patrick M Reed. Many-objective optimization and visual analytics reveal key trade-offs for london’s water supply. *Journal of Hydrology*, 531:1040–1053, 2015.
- [92] Erum Mehmood and Tayyaba Anees. Challenges and solutions for processing real-time big data stream: A systematic literature review. *IEEE Access*, 8:119123–119143, 2020.

- [93] Joao Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. Ensemble approaches for regression: A survey. *ACM Computing Surveys (CSUR)*, 45(1):10, 2012.
- [94] Mustafa Mirik, Jack E Norland, Robert L Crabtree, and Mario E Biondini. Hyperspectral one-meter-resolution remote sensing in Yellowstone National Park, Wyoming: I. Forage nutritional values. *Rangeland Ecology & Management*, 58(5):452–458, 2005.
- [95] Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.
- [96] Ruihui Mu and Xiaoqin Zeng. A review of deep learning research. *KSII Transactions on Internet and Information Systems (TIIS)*, 13(4):1738–1764, 2019.
- [97] David J Mulla. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering*, 114(4):358–371, 2013.
- [98] E.Y. Nakagawa, P.O. Antonino, F. Schnicke, R. Capilla, T. Kuhn, and P. Liggesmeyer. Industry 4.0 reference architectures: State of the art and future trends. *Computers and Industrial Engineering*, 156, 2021.
- [99] Kathryn B Newhart, Ryan W Holloway, Amanda S Hering, and Tzahi Y Cath. Data-driven performance analyses of wastewater treatment plants: A review. *Water research*, 157:498–513, 2019.
- [100] P. Nthutang and A. Telukdarie. Integration of Small and Medium Enterprises for Industry 4.0 in the South African Water Services Sector: A Case Study for Johannesburg Water. In *IEEE International Conference on Industrial Engineering and Engineering Management*, 2019.
- [101] F J Nutman and F M Roberts. Coffee leaf rust. *PANS Pest Articles & News Summaries*, 16(4):606–624, 1970.
- [102] Miguel Oliveira and Daniel Afonso. Industry Focused in Data Collection: How Industry 4.0 is Handled by Big Data. In *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, DSIT 2019, pages 12–18, New York, NY, USA, 2019. Association for Computing Machinery.
- [103] Darian Onchis. Observing damaged beams through their time–frequency extended signatures. *Signal Processing*, 96, Part A:16–20, 2014.
- [104] A. V. Oppenheim and R. W. Schaffer. Spectrogram Display of the Time-Dependent Fourier transform of Speech, 2009.
- [105] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab. The Continuous-Time Fourier Transform, 1996.



- [106] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab. The Discrete-Time Fourier Transform, 1996.
- [107] G Pahl, K Wallace, L T M Blessing, W Beitz, and F Bauert. *Engineering Design: A Systematic Approach*. Springer London, 2013.
- [108] M. Paiva, A. Vasconcelos, and B. Frago. Using enterprise architecture to model a reference architecture for industry 4.0. In *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems*, volume 2, pages 709–716, 2020.
- [109] Dorota Palka and Jolanta Ciukaj. Prospects for development movement in the industry concept 4.0. *Multidisciplinary Aspects of Production Engineering*, 2:315–326, 09 2019.
- [110] Hyunwoo Park, Marcus A Bellamy, and Rahul C Basole. Visual analytics for supply network management: System design and evaluation. *Decision Support Systems*, 91:89–102, 2016.
- [111] Eduardo H.M. Pena, Sylvio Barbon, Joel J.P.C. Rodrigues, and Mario Lemes Proenca. Anomaly detection using digital signature of network segment with adaptive ARIMA model and Paraconsistent Logic. In *Proceedings - International Symposium on Computers and Communications*, 2014.
- [112] Miguel Piamonte, Monica Huerta, Roger Clotet, John Padilla, Tito Vargas, and David Rivas. WSN Prototype for African Oil Palm Bud Rot Monitoring. In *International Conference of ICT for Adapting Agriculture to Climate Change*, pages 170–181. Springer, 2017.
- [113] W. Piao, C. Kim, S. Cho, H. Kim, M. Kim, and Y. Kim. Development of a protocol to optimize electric power consumption and life cycle environmental impacts for operation of wastewater treatment plant. *Environ Sci Pollut Res Int*, 23(24):25451–25466, Dec 2016.
- [114] Artzai Picon, Aitor Alvarez-Gila, Maximilian Seitz, Amaia Ortiz-Barredo, Jone Echazarra, and Alexander Johannes. Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. *Computers and Electronics in Agriculture*, 2019.
- [115] Barani R Priyanga and DrKANitha Kumari. A Survey on Anomaly Detection using Unsupervised Learning Techniques. *International Journal of Creative Research Thoughts (IJCRT)*, 6(2):2320–2882, 2018.
- [116] A. Puchalski. A technique for the vibration signal analysis in vehicle diagnostics. *Scopus*, 56:173–180, 2015.
- [117] Humberto Gutiérrez Pulido, Román De la Vara Salazar, Porfirio Gutiérrez González, Carlos Téllez Martínez, and María del Carmen Temblador Pérez. *Análisis y diseño de experimentos*. McGraw-Hill New York, NY, USA., 2012.

- [118] Julien Rabatel, Sandra Bringay, and Pascal Poncelet. Anomaly detection in monitoring sensor data for preventive maintenance. *Expert Systems with Applications*, 2011.
- [119] Annie Ibrahim Rana, Giovani Estrada, Marc Sole, and Victor Munteș. Anomaly Detection Guidelines for Data Streams in Big Data. In *Proceedings - 2016 3rd International Conference on Soft Computing and Machine Intelligence, ISCOMI 2016*, 2017.
- [120] Susmita Ray. A quick review of machine learning algorithms. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 35–39, 2019.
- [121] Gopinath Rebala, Ajay Ravi, and Sanjay Churiwala. *Machine Learning Definition and Basics*, pages 1–17. Springer International Publishing, Cham, 2019.
- [122] Luis Ribeiro and Mats Bjorkman. Transitioning from Standard Automation Solutions to Cyber-Physical Production Systems: An Assessment of Critical Conceptual and Technical Challenges. *IEEE Systems Journal*, 2018.
- [123] C A Rivillas, C A Serna, M A Cristancho, and A L Gaitan. La roya del cafeto en Colombia: Impacto manejo y costos del control. Technical report, Cenicafe, 2011.
- [124] Santiago Ruiz-Arenas, Zoltán Rusák, Ricardo Mejía-Gutiérrez, and Imre Horváth. Implementation of system operation modes for health management and failure prognosis in cyber-physical systems. *Sensors*, 20(8):2429, 2020.
- [125] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition, 2010.
- [126] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.
- [127] Dinesh Kumar Saini, Dikshika Ahir, and Amit Ganatra. Techniques and challenges in building intelligent systems: Anomaly detection in camera surveillance. In Suresh Chandra Satapathy and Swagatam Das, editors, *Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2*, pages 11–21, Cham, 2016. Springer International Publishing.
- [128] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 2001.
- [129] Erich Schubert, Jörg Sander, Martin Ester, Hans Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Transactions on Database Systems*, 42:1–21, 07 2017.

- [130] Klaus Schwab. *The Fourth Industrial Revolution*. Crown Publishing Group, USA, 2017.
- [131] K. Schweichhart. Reference Architectural Model Industrie 4.0 (RAMI 4.0) - An Introduction), 2016. [Online; accessed 31-Jan-2020].
- [132] G Sha, M Radzieński, M Cao, and W Ostachowicz. A novel method for single and multiple damage detection in beams using relative natural frequency changes. *Mechanical Systems and Signal Processing*, 132:335–352, 2019.
- [133] Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal*, 2016.
- [134] Pramila P. Shinde and Seema Shah. A review of machine learning and deep learning applications. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCCUBEA)*, pages 1–6, 2018.
- [135] Basilio Sierra, Elena Lazkano, Ekaitz Jauregi, and Itziar Irigoien. Histogram distance-based bayesian network structure learning: A supervised classification specific approach. *Decision Support Systems*, 48(1):180–190, 2009.
- [136] Jonathan A. Silva, Elaine R. Faria, Rodrigo C. Barros, Eduardo R. Hruschka, André C.P.L.F. De Carvalho, and João Gama. Data stream clustering: A survey. *ACM Computing Surveys*, 46(1), 2013.
- [137] J. J. Sinou. An Experimental Investigation of Condition Monitoring for Notched Rotors Through Transient Signals and Wavelet Transform. *Experimental Mechanics*, 49(5):683–695, 2009.
- [138] J.O. Smith. *Introduction to Digital Filters: With Audio Applications*. Music signal processing series. W3K, 2007.
- [139] Nan-Yao Su. Remote monitoring system for detecting termites, September 1998. US Patent 6,052,066.
- [140] Nan-Yao Su. Remote monitoring system for detecting termites, 2000.
- [141] Susanto B. Sulisty, W. L. Woo, S. S. Dlay, and Bin Gao. Building a Globally Optimized Computational Intelligent Image Processing Algorithm for On-Site Inference of Nitrogen in Plants. *IEEE Intelligent Systems*, 2018.
- [142] Susanto B. Sulisty, Wai Lok Woo, and S. S. Dlay. Regularized Neural Networks Fusion and Genetic Algorithm Based On-Field Nitrogen Status Estimation of Wheat Plants. *IEEE Transactions on Industrial Informatics*, 2017.
- [143] Susanto B. Sulisty, Di Wu, Wai Lok Woo, S. S. Dlay, and Bin Gao. Computational Deep Intelligence Vision Sensing for Nutrient Content Estimation in Agricultural Automation. *IEEE Transactions on Automation Science and Engineering*, 2018.

- [144] Dong Sun, Renfei Huang, Yuanzhe Chen, Yong Wang, Jia Zeng, Mingxuan Yuan, Ting-Chuen Pong, and Huamin Qu. Planningvis: A visual analytics approach to production planning in smart factories. *IEEE transactions on visualization and computer graphics*, 26(1):579–589, 2019.
- [145] Swee Chuan Tan, Kai Ming Ting, and Tony Fei Liu. Fast anomaly detection for streaming data. In *IJCAI International Joint Conference on Artificial Intelligence*, 2011.
- [146] K. W. Taylor, P. N. Burns, J. P. Woodcock, and P. T. Wells. Blood flow in deep abdominal and pelvic vessels: ultrasonic pulsed-Doppler analysis. *Radiology*, 154:487–493, 1985.
- [147] J.J. van Wijk. The value of visualization. In *VIS 05. IEEE Visualization, 2005.*, pages 79–86, 2005.
- [148] Juan Vanerio and Pedro Casas. Ensemble-learning approaches for network security and anomaly detection. In *Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks*, Big-DAMA '17, page 1–6, New York, NY, USA, 2017. Association for Computing Machinery.
- [149] David Velásquez, Santiago Perez, Ricardo Mejia-Gutierrez, and Alejandro Velásquez-López. Crack detection method in transport of hygroscopic particulate compressed material. *International Journal of Mechanical & Mechatronics Engineering*, 20:26–33, 04 2020.
- [150] Vincent Vercauysen, Wannes Meert, Gust Verbruggen, Koen Maes, Ruben Baumer, and Jesse Davis. Semi-Supervised Anomaly Detection with an Application to Water Analytics. In *Proceedings - IEEE International Conference on Data Mining, ICDM, 2018*.
- [151] S. Wang, G. Cai, Z. Zhu, W. Huang, and X. Zhang. Transient signal analysis based on Levenberg-Marquardt method for fault feature extraction of rotating machines. *Scopus*, 54:16–40, 2015.
- [152] S. Webb, K. Peters, M. A. Zikry, S. Chadderdon, S. Nikola, R. Selfridge, and S. Schultz. Full-Spectral Interrogation of Fiber Bragg Grating Sensors Exposed to Steady-State Vibration. *Experimental Mechanics*, 53(4):513–530, 2013.
- [153] Weiliang Wu, Wenzhong Qu, Li Xiao, and Daniel J Inman. Detection and localization of fatigue crack with nonlinear instantaneous baseline. *Journal of Intelligent Material Systems and Structures*, 27(12):1577–1583, 2016.
- [154] Wenchao Wu, Yixian Zheng, Kaiyuan Chen, Xiangyu Wang, and Nan Cao. A visual analytics approach for equipment condition monitoring in smart factories of process industry. In *2018 IEEE Pacific Visualization Symposium (PacificVis)*, pages 140–149. IEEE, 2018.

- 
- [155] Min Xu, Jeanne M. David, and Suk Hi Kim. The fourth industrial revolution: Opportunities and challenges. *International Journal of Financial Research*, 2018.
- [156] Ruqiang Yan and Robert X. Gao. Multi-scale enveloping spectrogram for vibration analysis in bearing defect diagnosis. *Tribology International*, 42(2):293–302, 2009.
- [157] Ye Yuan, Shouzheng Li, Xingjian Zhang, and Jianguo Sun. A comparative analysis of svm, naive bayes and gbdt for data faults detection in wsns. In *2018 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*, pages 394–399, 2018.
- [158] Xiaojin Zhu and Andrew B. Goldberg. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009.

Part II  
Appended Papers



---

## Summary of the appended papers

---

This section provides a brief summary of the six papers that have been appended to this document. It includes their title, primary objective, methodology, and findings. Finally, the papers are attached.

### 11.1 Paper 1

**Title:** A Novel Architecture Definition for AI-driven Industry 4.0 Applications.

**Conference Paper:** Proceedings - 2022 11Th International Conference on Industrial Technology and Management (ICITM), IEEE Xplore.

**Year:** 2022.

**Objective:** To develop a novel generic architecture that allows to design and develop AI-driven Industry 4.0 systems.

**Methodology:** The methodology proposed in this paper is based on the analysis of three different case studies of industrial projects that improved or created a new product/service by using AI-driven Industry 4.0 technologies.

**Findings:** A novel software and hardware architecture for AI-driven Industry 4.0 systems was developed based on real-world use cases. The architecture was created by using common elements of different architectures presented in the state-of-the-art and by analysing three case studies of real industrial projects. It includes a physical layer, an embedded system layer and an IoT cloud layer with which it is possible to clearly organise all components of a 4IR system. This architecture provides a detailed



view and practical guideline for users on how to implement 4IR systems by embedding relevant hardware and software components.

## 11.2 Paper 2

**Title:** A Cyber-Physical Data Collection System Integrating Remote Sensing and Wireless Sensor Networks for Coffee Leaf Rust Diagnosis.

**Journal:** Sensors, MPDI.

**Year:** 2021.

**Objective:** Data acquisition combined with data preparation is the first part of a Machine Learning process. In this paper the main objective is to design and implement a cyber-physical data collection system that integrates Remote Sensing and Wireless Sensor Networks, which can be used to train a Machine-Learning model to diagnose Coffee Leaf Rust (CLR).

**Methodology:** Previous research was helpful to design a cyber-physical data collection system that could integrate both methods towards the CLR diagnosis. Applying the concepts and following the recommendations found in the state-of-the-art, it is possible to create a system capable of acquiring and remotely storing reliable data from diverse sources. The goal of such cyber-physical systems (CPS) is the characterisation of a test bench coffee-crop regarding the changes induced by the disease at hand. The cyber-physical data collection system was designed following the Pahl and Beitz methodology.

**Findings:** A cyber-physical data-collection system was developed, by integrating Remote Sensing and Wireless Sensor Networks, to gather data, during the development of the CLR, on a test bench coffee-crop. The system is capable of automatically collecting, structuring, and locally & remotely storing reliable multi-type data from different field sensors, Red-Green-Blue (RGB) and multi-spectral cameras (RE and RGN). In addition, a data-visualization dashboard was implemented to monitor the data-collection routines in real-time. The operation of the data collection system allowed to create a three-month size dataset that can be used to train CLR diagnosis machine learning models. This result validates that the designed system can collect, store, and transfer reliable data of a test bench coffee-crop towards CLR diagnosis.

## 11.3 Paper 3

**Title:** A Method for Detecting Coffee Leaf Rust through Wireless Sensor Networks, Remote Sensing, and Deep Learning: Case Study of the Caturra Variety in Colombia.

**Journal:** Applied Sciences, MDPI.

**Year:** 2020.

**Objective:** Machine Learning algorithms commonly require a significant amount of data to perform an optimal model, and by ensembling, it is possible to combine the output of several models into one. This project aimed to develop a coffee leaf rust stage classification (CLR) model by integrating Remote Sensing (RS), Wireless Sensor Networks, and Deep Learning techniques.

**Methodology:** This paper proposes a method to do a machine learning ensembling by integrating WSN, RS and deep learning techniques for detecting the Coffee Leaf Rust (which is an anomaly to a coffee plant). The ensembling consists of a weighted average by using the  $F_1$ -score of each model.

**Findings:** An ensemble model comprising the integration of WSN, RS, and Deep Learning was implemented to detect the CLR with an F1 score of 0.775. The analysis of the results revealed a p-value of 0.231, which indicated that the difference between the disease diagnosis made employing a visual inspection and the proposed technological integration was not statistically significant. The above shows that both methods were significantly similar in diagnosing the disease.

## 11.4 Paper 4

**Title:** A Hybrid Machine-Learning Ensemble for Anomaly Detection in Real-Time Industry 4.0 Systems.

**Journal:** IEEE Access, IEEE.

**Year:** 2022

**Objective:** The detection of faults and anomalies in real-time industrial systems is a challenge due to the difficulty of sufficiently covering an industrial system's complexity. This paper proposes to develop a hybrid machine-learning ensemble real-time anomaly-detection pipeline that combines three Machine-Learning models -a Local Outlier Factor, a One-Class Support Vector Machine, and an Autoencoder-, through a weighted average, to improve anomaly detection.

**Methodology:** The proposed ML hybrid pipeline for real-time anomaly detection, consists of two stages: i) the Manufacturing stage and ii) the Operation stage. The Manufacturing stage takes its name from the manufacturing process of an industrial machine. At this stage, an ML model is trained on machines' quality control process

data to validate whether the machine meets its design standards or not. The operation stage or pipeline refers to the phase when the machine is already running in production; in terms of a classical ML pipeline, it represents the deployment phase. It uses the pre-trained model from the manufacturing stage to detect anomalies in real-time. The ML models inside this pipeline combines three Machine-Learning models -a Local Outlier Factor, a One-Class Support Vector Machine, and an Autoencoder-, through a weighted average to develop a ensemble model.

**Findings:** This research work has developed and presented a Hybrid Machine-Learning Ensemble for Anomaly Detection for a Real-Time Industry 4.0 System. This ensemble consists of implementing two stages inspired by a standard industrial system: i) A Manufacturing Stage and ii) An Operation Stage. Up to our knowledge, there are no other ML methods that consider these industrial stages. The ensemble system was tested on three machines, presenting an increased  $F_1$ -score value and AUC concerning individual ML submodels (LOF, OCSVM, and Autoencoder). The ensemble model for Machine A presented a  $F_1$ -score value of 0.904 for anomalies (-1), a  $F_1$ -score value of 0.944 for normal data (1), and an AUC value of 0.913; the ensemble model for Machine B presented a  $F_1$ -score value of 0.890 for anomalies (-1), a  $F_1$ -score value of 0.946 for normal data (1), and an AUC value of 0.905; finally, the ensemble model for Machine C presented a  $F_1$ -score value of 0.887 for anomalies (-1), a  $F_1$ -score value of 0.889 for normal data (1), and an AUC value of 0.897.

## 11.5 Paper 5

**Title:** Crack Detection Method in Transport of Hygroscopic Particulate Compressed Material.

**Journal:** International Journal of Mechanical and Mechatronics Engineering.

**Year:** 2020

**Objective:** To develop a crack detection method in transport of hygroscopic particulate compressed materials by using frequency and spectral analysis.

**Methodology:** This article proposes a methodology that first verifies if a product sample will resist the transportation conditions, simulating them through hardware and software. The hardware consists of a vibrations test bench, which first makes the sample oscillate at a particular frequency and with a simulated spring pressure. The vibrations' data of the sample is then acquired using an accelerometer through a data acquisition device, which reports the information in a computer database. The software component consists of a developed spectrogram post-processing algorithm that checks the stored database and shows if the sample failed the test (detected crack) or if it passed the simulated transportation conditions.

**Findings:** A method for detecting failures in products being transported was implemented, using frequency analysis to verify if a crack occurred during the transportation of a given compressed product. The method considers vehicle vibrations and all possible interactions between the compressed material, type of packaging, and the vertical load applied by the compressed product over the sample (pile of the same product). A crack can be detected in a compressed hygroscopic particulate material by finding the time in the spectrogram when new frequency components, different from the main oscillating frequency, start to appear. These new frequency components, when they are detected, indicate that some detached particles from the primary sample started to oscillate around the material. These results indicate that frequency analysis can be used to detect anomalies as an alternative way to Machine Learning.

## 11.6 Paper 6

**Title:** EDAR 4.0: Visual-Analytics for Waste Water Management.

**Journal:** Sent to IEEE Transactions on Industrial Informatics (Under Review).

**Year:** 2023

**Objective:** Wastewater treatment plants (WWTPs) have large amounts of data from their sensors and from the tests performed every day in relation to water quality variables. Analysing all these variables to make decisions presents a major research challenge, where the use of Machine Learning and Visual Analytics tools can facilitate this task. For this reason this research aims to develop a Visual Analytics tool for Wastewater Treatment Plants.

**Methodology:** The methodology followed in this article is based on Avruskin [8], who proposes three steps for successful Data Analysis and Data Mining: i) Data collection and acquisition, where information on the target variables is obtained and measured; ii) Data management and validation, which consists of verifying the accuracy and quality of the data source before it is used; and iii) Data visualisation, where the data is graphed and represented.

**Findings:** This paper presents a visual-analytics-based platform for WWTP, called EDAR 4.0. Intuitive visualisations have great potential for supporting decision-making during the operation and management of WWTPs. The proposed tool allows users to identify relationships between data through simple data inspection. The developed tool allows WWTP operators to perform simulations and optimisations without risking real site testing. This tool has been validated with WWTP domain experts, showing that it can provide an additional source of information for WWTP management.



## CHAPTER 12

---

Appended papers

---



# A Novel Architecture Definition for AI-Driven Industry 4.0 Applications

David Velásquez\*,  
RID on Information Technologies and  
Communications Research Group,  
Universidad EAFIT, Colombia &  
Department of Data Intelligence for  
Energy and Industrial Processes,  
Vicomech Foundation,  
Spain  
dvelas25@eafit.edu.co

Mauricio Toro,  
RID on Information Technologies and  
Communications Research Group,  
Universidad EAFIT,  
Colombia  
mtorobe@eafit.edu.co

Jan L. Bruse,  
Department of Data Intelligence for  
Energy and Industrial Processes,  
Vicomech Foundation,  
Spain  
jbruse@vicomech.org

Xabier Oregui,  
Department of Data Intelligence for  
Energy and Industrial Processes,  
Vicomech Foundation,  
Spain  
xoregui@vicomech.org

Mikel Maiza,  
Department of Data Intelligence for  
Energy and Industrial Processes,  
Vicomech Foundation,  
Spain  
mmaiza@vicomech.org

Basilio Sierra,  
Department of Computer Science and  
Artificial Intelligence, University of  
Basque Country (UPV/EHU),  
Spain  
b.sierra@ehu.es

**Abstract**—The adoption of new technologies such as the Internet of Things and Big Data has become one of the main challenges in the era of Industry 4.0. Industrial 4.0 systems can provide increased reconfigurability and flexibility, in particular regarding Cyber-Physical Systems that integrate advanced artificial intelligence (AI) and fast-communication systems to provide real-time decisions. However, adoption of these new technologies in real-world applications is typically hindered by the complexity introduced by a multitude of legacy and state-of-the-art hardware and software components that need to be connected and need to communicate with each other. Consulting previously published approaches and analysing three real industrial use cases, this paper proposes a novel software-and-hardware architecture design for AI-driven Industrial 4.0 systems to facilitate the transition towards such smart-connected systems.

**Keywords**—industry 4.0, internet of things, architecture, cyber-physical system, artificial intelligence, system design

## I. INTRODUCTION

The fourth industrial revolution (4IR) sets new challenges for traditional industrial processes: to improve existing or create new processes that efficiently use novel technologies and take full advantage of their potential. In an increasingly competitive market, 4IR can be considered as a disruptive innovation that positively impacts different industrial sectors by integrating new enabling technologies. Examples of these technologies are 3D printing, Internet of Things (IoT), Cyber-Physical Systems (CPS), Artificial Intelligence (AI), Big Data, Robotics, Nanotechnology, and Quantum Computing [1], [2].

With the improvement of internet speed, coverage and bandwidth, 4IR systems can take advantage of cloud

computing [3] to process large volumes of data. These large volumes of data come from different types of sources such as sensors, supervisory control and data acquisition (SCADA) systems or third-party data sources (e.g., weather stations). After data has been acquired, AI algorithms process these large volumes of data to provide additional knowledge that may optimise processes and increase profitability. However, some types of applications require real-time processing, such as embedded system processing or edge computing technologies [4] to provide faster responses to the 4IR system.

Common contexts of 4IR applications include manufacturing, where massive data can be analysed in real-time to improve factory operations and production, thus reducing machine down times, which finally improves the product quality [5]. Another application context is the smart water-management industry, which uses process digitalisation and automation to gather useful information from water processing plants and external sources (e.g., weather information). This data is then processed through AI to optimise process efficiency, resource savings, and quality of results [6]–[8].

A key application for AI-driven 4IR applications is the field of (predictive) maintenance. In this sense, 4IR systems can provide process monitoring where the state of a system can be determined through three stages: detection, diagnosis, and prognosis.

The *detection* stage's main objective is to infer an anomaly accurately (detection strength) as soon as it happens (detection speed). Methods for such a task include statistical normal-operation condition models (NOC), which do not require

---

\* Corresponding author.



information about the process structure but instead estimate some of the model parameters through data processing. Some univariate examples of these NOC detection models are Shewhart, Exponentially-Weighted Moving Average (EWMA) and Cumulative Sum (CUSUM) control charts based on a univariate Gaussian model. Multivariate examples include the Hotelling's  $T^2$  chart, the Multivariate Exponentially Weighted Moving Average chart (MEWMA), Multivariate Cumulative Sum chart (MCUSUM), high-dimensional Principal Component Analysis (PCA), and Partial Least Squares (PLS) [9].

The *diagnosis* stage takes place when an anomaly is detected at the detection stage. Its main objective is to find possible causes for that anomaly or system failure, also known as “failure identification”. For this task, two existing structured approaches are used: i) knowledge-based and ii) data-driven. Knowledge-based approaches (e.g., Causal Maps, bond graphs, signed digraphs, parity relations, Bayesian Networks among others) make use of pre-created information on the whole causal connectivity of the process to find which part of the system was the cause of the anomaly or failure. Data-driven approaches (e.g., machine learning, one-class classifiers, multi-class classifiers, density-based methods, to name a few) extract data from logs and database records of the system operation to correctly classify features that caused the anomaly or failure [9].

Finally, the prognosis is the most challenging stage, which requires a stable operation of the two previous stages (detection and diagnosis). This stage focuses on predictions of future faults, called “fault prognosis”. Thus, prognosis can provide better planning of maintenance, which can optimise 4IR systems process performance by minimising production losses and by providing a more secure working environment for both workers and all the process equipment, hence maintaining Equipment Health [9].

Hence, two significant problems arise when designing a novel, AI-driven 4IR system: i) How to correctly design the system from scratch [10], [11] and ii) How to improve existing legacy systems to integrate smart-capable and connectivity layers [12].

The first problem is related to the process of 4IR system design, where depending on the requirements for the design, all the parts of the system must be correctly selected and interfaced to satisfy a set of characteristics that these novel systems must have [10], [13]–[15]. These characteristics are: Customization, where the system is designed to adjust to product families instead of to a single product; Convertibility, to facilitate functional changes; Scalability, to adapt to different process demands by increasing or reducing the usage of software and hardware resources; Modularity, where each system component is considered as a module and provides a specific function; Integrability, to allow quick integration of future system upgrades from software (Cyber) or hardware (Physical); and Diagnosability, to react to disturbances and detect any possible anomaly that may be occurring. Moreover, many of these characteristics of 4IR systems are part of the technical perspective of sustainable systems [16]. Unfortunately, in a systematic mapping study on sustainable

4IR systems, authors found that the specification of these characteristics is not often considered, and most of the designs do not consider the physical layer [16].

The second problem is related to the fact that existing industry systems are typically isolated and with limited connection to external systems. These industry systems are usually driven by Programmable Logic Controllers (PLCs), which are only accessible through a local industry network. These legacy systems are inflexible when upgrading or expanding their functionalities in a ubiquitous manner [17]. While current research involves incorporating the Internet of Things and self-adaptation to such systems, most of the case studies are purely theoretical and have not been tested in an industrial environment [18], [19].

To address these two problems, 4IR system architectures that define all their components have been proposed in the state-of-the-art by different authors, comprising technologies such as IoT, CPS, and smart-systems: Ganti et al. [20] presented a Mobile Crowdsensing (MCS) technology review, commonly used for environmental, infrastructure, and social applications. Functional architecture is presented in their review and shows how data coming from different contexts can be interconnected and provide useful information to end-users. Their architecture consists of seven parts. First, there is a “context” with all the raw data that can be sensed through “sensors”. Then, a “localized analytics” module processes and analyses these raw data (e.g., using embedded system processors or edge computing technologies [4]) to provide a summary of data “privately” to the local user in a real-time manner. Then, massive data coming from different contexts are sent to a cloud where a “Back-end server aggregation module” can store it, and, eventually, send it to an “aggregate analytics module”. The analytics module can then analyse all the data using advanced AI-driven algorithms and finally show it to remote users through an “App”.

Bagheri et al. [21], and Lee et al. [22] proposed a unified framework that integrates CPS in manufacturing. This unified framework architecture contains a 5-level definition (5C framework) that allows CPS manufacturing design and deployment from data acquisition to data analysis and final value creation. The first level is named “Smart Connection Level” which acquires accurate and reliable data from machines and components through sensors obtained from a process controller or through the enterprise manufacturing system (e.g., an enterprise resource planning ERP system). The second level is the “Data-to-Information Conversion Level”, which is in charge of inferring valuable information from the data and adding self-awareness to the 4IR CPS system. The third level, named “Cyber Level”, acts as a central information hub, gathering massive amounts of data, which are then roughly analysed to acquire additional information. This new information provides insights into the status of the 4IR system and possible predictions of its future behaviour, giving a self-comparison ability. The fourth level is called the “Cognition Level”, where a 4IR CPS system generates knowledge of a particular part of the system or a process for an expert user, presented using visual analytics and infographics tools. Then, through this knowledge, expert users can make knowledge-guided decisions. Finally, the fifth level is the “Configuration

level”, whose main objective is to provide a feedback loop from cyberspace to the physical space, acting as a supervisory control making machine or process changes depending on all the previous information knowledge. This level adds self-configure and self-adaptive quality features to 4IR systems. Fig. 1 provides a detailed view of the manufacturing CPS 5C framework architecture.

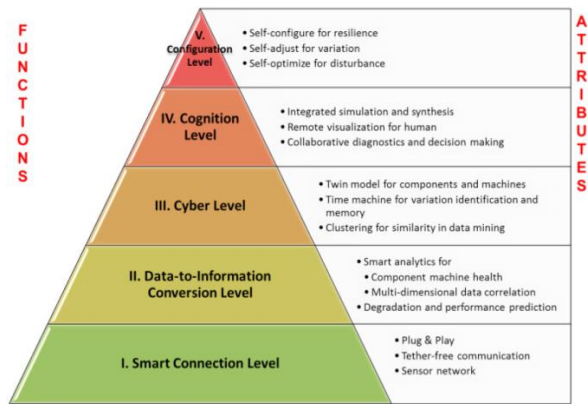


Fig. 1. 5C framework architecture [21].

Blonda et al. [6] propose an IoT Middleware architecture to expose its functionalities as a set of Cloud-supported RESTful APIs. Three layers compose Blonda et al.’s architecture: Users, Middleware, and Physics layer. The Middleware architecture is divided into three sublayers: application, network, and security. According to Blonda et al., security in IoT systems can be defined using six properties: confidentiality, integrity, availability, identification and authentication, privacy, and trust. Fig. 2 provides a detailed view of their architecture.

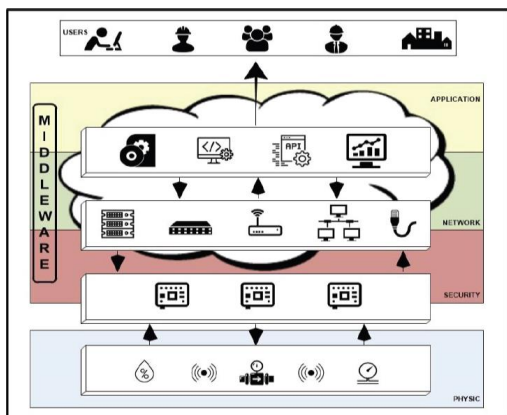


Fig. 2. Middleware architecture for IoT systems [6].

On the other hand, there is the Reference Architectural Model for Industrie 4.0 (RAMI 4.0) [23], which is defined in three dimensions. The first dimension is the life cycle value stream, defined in standard IEC 62890, composed of the following phases: Type and Instance. The second dimension are hierarchy levels, defined in standard IEC 62264 and IEC

61512. The levels are the following: product, field device, control device, station, work center, enterprise, connected world. Finally, the third dimension are different layers, similar to previous architectures: business, functional, information, communication, integration, and asset. Fig. 3 provides a compact view of the architecture.

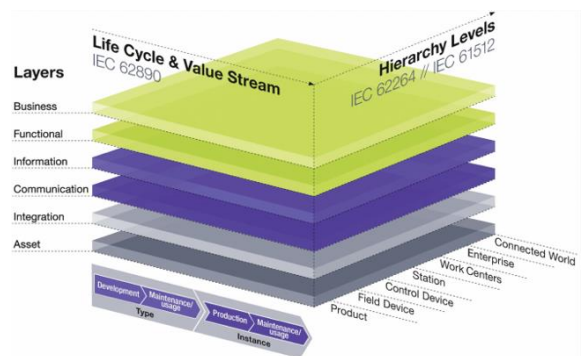


Fig. 3. Reference Architectural Model for Industrie 4.0 (RAMI 4.0) [23].

Paiva et al. [24] proposes a hybrid reference architecture called RAMI 4.0 EA that integrates RAMI 4.0 and Enterprise Architecture (EA). The proposed reference architecture allows empowering RAMI 4.0 with elements such as EA principles, applications, technology, and organizational processes in a visual and easy-to-understand way, allowing enterprises’ better adoption in 4IR projects.

Finally, the Industrial Internet Reference Architecture (IIRA) was created by the Industrial Internet Consortium (IIC) [25]. It is application domain-independent, and its development was guided by industry. IIRA is focused on the functionalities required by industry, specifically in prognostics, optimisation, operation, business, analytics, and the monitoring and control of devices.

Some approaches for defining components integrated into a 4IR system are a mobile crowdsensing functional architecture, a 5C framework architecture, and an IoT middleware architecture. They all have in common that there must be a physical layer, a middleware, and a layer for user interaction. The middle layer’s organisation is presented differently in each architecture, but they all underline the importance of security and data analytics. Additionally, there is a lack of guidelines explaining how to organise the components, their connections and their interfaces in practical terms to get a detailed architecture of the system and end up with a working system [26].

The methodology proposed here addresses the previously identified problems by taking into account the best state-of-the-art practices for designing from zero a 4IR system architecture and by analysing different industrial case studies to give a comprehensive understanding of the given problem. In this way, all components around a smart-connected system that provide direct feedback to the user about the status of a process or a system can be defined.

This article is divided into four (4) sections. First, the introduction section described the concepts and importance of 4IR systems. It also includes a review of the current state-of-the-art highlighting different approaches and existing methods for defining Industry 4.0 architectures. Then, a case study analysis of three different architectures from real industrial implementations is presented in section 2, which is used to define our proposed novel methodology for defining Industry 4.0 architectures. The results section describes the details of the proposed generic, extensible, and flexible architecture. Finally, conclusions are drawn.

## II. PROPOSED METHOD

This section describes the proposed method, which is based on the analysis of different case studies of industrial projects that improved or created a new product/service by using AI-driven Industry 4.0 technologies. This section will first describe three real industrial case studies. Their processes and technological architectures (hardware and software components) are described and compared, followed by the identification and selection of components that we considered to be the basis of our proposed, generic 4IR system architecture.

### A. Smart-Water Case Study: Industrial Wastewater Treatment Plant “La Cartuja / EDAR 4.0 Project”

EDAR 4.0 is a research project aiming to develop a set of tools for optimising the operation and in particular the energy management of wastewater treatment plants (WWTPs). Different types of organisations such as water and energy engineering companies, process automation companies, WWTP equipment manufacturing companies, research centres, and universities participate in the project, all collaboratively working on different aspects of the project, finally aiming to develop a cloud-based, web platform integrating a complete set of tools for supporting an intelligent operation of WWTPs.

The project's basis consists of plant-wide data acquisition of all the processes comprising a WWTP. These processes can be classified into three principal, standard sub-processes: i) the influent process, mainly representing the input of influent water and its pre- and primary treatment, usually performed in a primary settling or sedimentation tank; ii) the biological treatment process, which is the central part of the so-called secondary treatment and represents the primary wastewater treatment process of the plant driven by different types of bacteria and protozoa, which can be complemented by additional, chemical treatments, and; iii) the effluent process, which mainly represents the output of the effluent water, either directly to the receiving waters or through a secondary settling or sedimentation tank, which is also considered as part of the secondary treatment of the plant. A tertiary treatment process consisting of additional, advanced water purification treatments aimed at specific water uses such as water reuse can exist but is optional and not so frequent. In this project, sub-processes i) to iii) of the full scale WWTP are addressed.

The processes of a WWTP in general and sub-processes i) to iii) in particular, are typically controlled by one or several Programmable Logic Controllers (PLC), integrated with different types of sensors and actuators. All the control information is then locally displayed through Human to

Machine Interfaces (HMI), generally embedded within a Supervisory Control And Data Acquisition (SCADA) system. All plant information is usually shared through an industrial protocol-based local area network (LAN).

The above represents the basis of a typical WWTP, ICT architecture. In EDAR 4.0, this is extended to a 4IR system architecture by setting an additional, cloud-based IoT infrastructure that can be reached through the internet, thus the overall WWTP, ICT infrastructure has to have (a secure) access to it. Several services such as multiple plants, cloud-based IoT data acquisition and data storage, information monitoring (visualisation), data analysis and associated services such as Visual Analytics, plant simulation, and plant optimisation through machine learning (ML), are integrated into such a cloud-based ICT infrastructure.

A specific example of how to access the above cloud IoT infrastructure and associated services could be through the HTTP REST protocol. A specific example of a data analysis service could be to classify different types of water quality and predict (forecast) the evolution of water quality over time.

Eventually, with the above cloud IoT platform running, WWTP data can be displayed on a web page, where water quality analyses and others can be run and monitored by remote users. Fig. 4 details a view of the EDAR 4.0, 4IR system architecture.

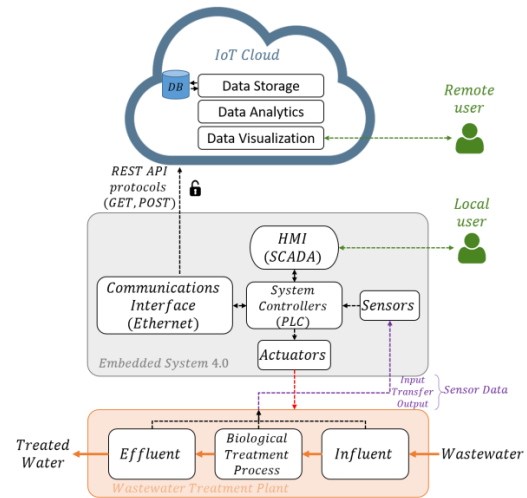


Fig. 4. EDAR 4.0 architecture [Original Work].

### B. Industrial Quality Testbench Case Study: Rotary Pneumatic Machines Company “MAPNER / EDAR 4.0”

MAPNER is an industrial company that manufactures rotary machines for various applications such as wastewater treatment and power generation. Once the manufacturing process of the rotary machines has finished, every machine is taken to a quality control process performed on a testbench where the machines are subjected to a set of tests in stationary working conditions, in an isolated room. The outputs of the tests are then compared to some expected results described by

the manufacturing order, in order to guarantee an adequate quality and performance rate of the final product (the machine).

However, that process is typically highly manual: after leaving the machine on for some time until it reaches its stationary operation region, the operator goes through a set of GUI elements of a computer program. It shows the data measured by sensors and allows manually introducing such data to a database, so that subsequent calculations of physical magnitudes such as flow rate and power can be performed to generate a quality report of the machine.

This use case is a good example of a classic manufacturing process digitalisation project where the process evolves from a view-only data management system to an automatic, real-time data acquisition and storage system, which not only improves the existing testing process (the machine's performance can be analysed continuously instead of via a single-instant, manual data acquisition system, which can hardly reflect the overall condition of the rotary machine), but also allows stepping forward towards a data-analysis-based machine performance study that may facilitate identifying, predicting and preventing problems for the manufactured products.

The first step of adapting MAPNER's testbench to a 4IR platform was automating the acquisition of data corresponding to measurements as provided by sensors attached to the machine during the testing process as well as by some environmental sensors in charge of measuring relative humidity and temperature. As usual in many manufacturing processes, every sensor or measuring device has its own communication protocol for providing the information. Multiple protocol systems are usually managed by using gateways that unify the information and translate it into a standard protocol. These gateways can be independent hardware devices or software modules designed to do so. In this use case, a software gateway was implemented to gather all the information on a single, Python-based daemon (or "Python gateway") in order to be able to send the data to the data storage layer. In order to do so, the data had to be converted (unified) to a common, standard communications interface: Ethernet. Machine sensors are connected to a Siemens PLC that exposes the information by means of the OPC-UA protocol. The humidity and temperature information, as well as the information provided by a set of electrical network analysers (current, voltage, power, power factor, etc.) are exposed by means of the MODBUS TCP protocol, and transferred to Ethernet by means of a MODBUS-MODBUS TCP hardware gateway.

The Python gateway did not only made the acquisition possible, but it also helped fixing a critical aspect related to acquiring data from multiple sources - thanks to the gateway, incoming data with different sampling frequencies can be homogenised by applying data synchronisation or re-sampling algorithms prior to analysis.

Once the data coming from the different sources is gathered and unified, the information is deployed to an internal server database, where it can be further processed and analysed in order to create and show to operators enriched information (such as machine status information) throughout the tests, in real-time. The data storage is running within a secure LAN. In

addition, a daily backup of the information managed by the data storage layer is configured. The global architecture of the entire system is shown in Fig. 5.

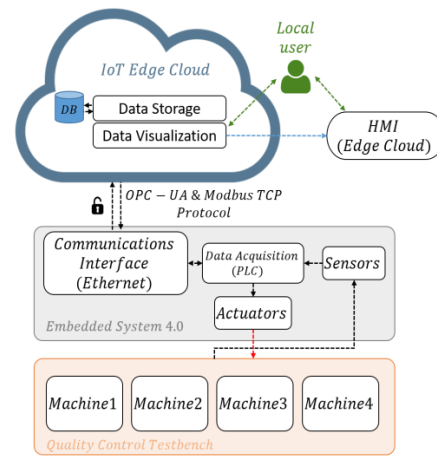


Fig. 5. MAPNER Testbench architecture [Original Work].

### C. Smart IoT Embedded System Case Study: Rotary Pneumatic Machines Company "MAPNER / SISTELIA Project"

The third case study consists of an architecture defined for remotely managing data related to blower machines manufactured by MAPNER that are installed in various locations across the world. With the arrival of the digital transformation and the 4IR, MAPNER clearly saw the opportunity for providing their machines with a greater degree of ubiquity, especially in terms of smart and predictive maintenance, using technologies such as IoT and AI. In this sense, MAPNER has participated in several funded R&D projects and collaborated with research agents. One of these projects is SISTELIA, which translates to "Intelligent Services for Industrial Blowers based on Digitalization Technologies and Artificial Intelligence". SISTELIA's main objective is to design and implement a cloud-based data management platform based on a 4.0 embedded system called "MAPNER Panel Control" (MPC), designed for the acquisition, analysis, and visualisation of data and enriched information coming from machines that are operating worldwide, in real-time. The MPC includes an ad hoc Human Machine Interface (HMI) which gathers data from an integrated, real-time data acquisition (DAQ) system that is integrated with sensors and provides real-time information to maintenance operators both directly on the machine (local visualisation) and through the cloud (remote visualisation), via a 4G network Communications Interface. SISTELIA's architecture consists of three parts: i) the physical blowing machine, ii) a 4IR embedded system, and iii) a cloud data management platform. The physical blowing machine is the core process of the whole system, which operates independently of the architecture. The 4IR embedded system contains multiple sensors to gather operational data (e.g., temperature, speed, pressure, and vibrations) from the physical blowing machine, which is locally stored and processed in a real-time DAQ system.

Furthermore, the locally stored and processed data is also locally displayed to users on an integrated, tactile HMI display, where the user can program maintenance operations and configure and resolve alarms. In addition, these data are sent to an IoT cloud platform via 4G for remote storage, real-time visualisation, and data analysis. Finally, the cloud data can be monitored by remote users. This architecture is detailed in Fig. 6.

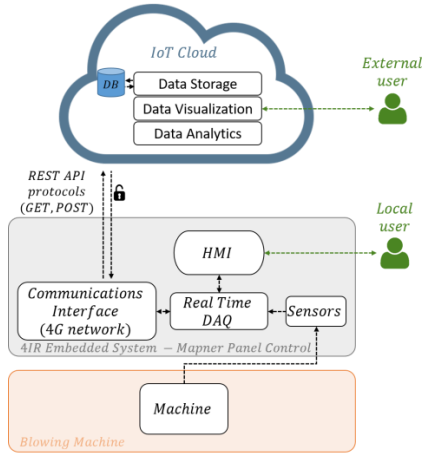


Fig. 6. SISTELIA architecture [Original Work].

### III. RESULTS

Results of this work consist of a generic 4IR architecture that aims to serve as a template for new AI-driven Industry 4.0 projects. This architecture has been built from hardware and software components identified in the use cases presented in the previous section as well as in the state-of-the-art.

The proposed generic architecture presented in Fig. 7 includes three levels: i) the physical layer, ii) the layer of an embedded 4IR system, and iii) the IoT cloud layer. The physical layer relates to the process itself; for example, a machine that executes a task.

The 4IR embedded system layer may include different sub-layers of Perception and Control, where everything related to actuators and sensors can be found; Data Acquisition and Processing, which involves the different micro controller and PLC units with their respective internal/external storage systems for local data persistence; the Local Visualisation sublayer, which includes the different HMI interfaces for the visual and control interaction between the local user (who could be a supervisor or an operator on site) and the machine; the Communications sublayer, which addresses all the local communication interface systems such as RS232, RS485, Modbus, Profibus, and the global ones (for the new 4IR systems) through TCP/IP protocol by Ethernet, WiFi and 3G/4G networks, allowing to connect to the IoT cloud.

The IoT Cloud layer incorporates four sub-layers. The first one is the Security and Data Exchange sublayer, which establishes a secure connection between the Embedded 4IR System and other external information sources (External Data)

through WebAPIs. The latter, for example, can use WebSockets for real-time connections, HTTP REST protocol for on-demand requests, Mosquitto Transfer Protocol (MQTT) for IoT connections, OPC-UA for industrial data connections, among others.

The second sub-layer is related to Data Storage, where relational (e.g., SQL) and non-relational databases are used. This stored data can then be retrieved to perform different analytical and visualisation operations, for example.

The third sub-layer is called Analysis, in which AI tools perform advanced processing operations. These operations create enriched information, thus adding a greater degree of knowledge about the process to allow, for example, identifying failures in a predictive manner.

Finally, the fourth sub-layer includes the Remote Visualisation, where dashboards are usually shown with the received data and additional graphics derived from the analysis process (e.g., Visual Analytics) for supervision by External Users.

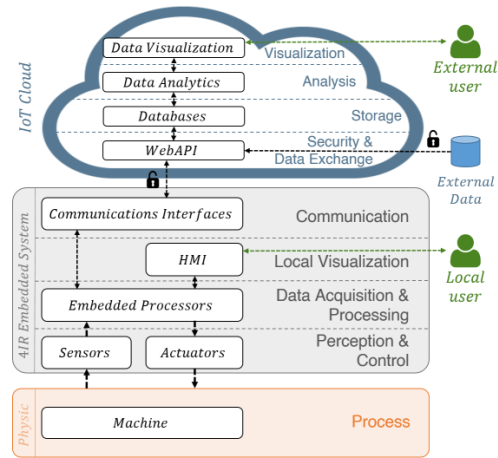


Fig. 7. 4IR generic architecture (contribution).

### IV. CONCLUSIONS

Industry 4.0 systems including AI applications are generally complex and difficult to understand as they consist of so many different HW and SW components. Typically, they include state-of-the-art and legacy technologies. Hence, guidelines are missing explaining how such a problem could be tackled in practical terms and in which way components and their connections and interfaces could be organised to have a full-sight understanding of the system and end up with a working system. In this sense, a novel software and hardware architecture for AI-driven Industry 4.0 systems was developed based on real-world use cases. The architecture was created by using common elements of different architectures presented in the state-of-the-art and by analysing three case studies of real industrial projects. It includes a physical layer, an embedded system layer and an IoT cloud layer with which it is possible to clearly organise all components of a 4IR system. This architecture provides a detailed view and practical guideline

for users on how to implement 4IR systems by embedding relevant hardware and software components.

Future work may include a methodology to generate a specific architecture based on both user requirements and the above generic architecture for a new 4IR project. Additionally, other case studies can be analysed to extend this architecture.

#### ACKNOWLEDGMENT

We would like to thank Vicomtech Foundation for providing the time and resources for this work. We also want to thank EAFIT University for sponsoring the research scholarship for the PhD.

#### REFERENCES

- [1] M. Xu, J. M. David, and S. H. Kim, "The fourth industrial revolution: Opportunities and challenges," *Int. J. Financ. Res.*, 2018, doi: 10.5430/ijfr.v9n2p90.
- [2] M. H. Lee *et al.*, "How to respond to the Fourth Industrial Revolution, or the second information technology revolution? Dynamic new combinations between technology, market, and society through open innovation," *J. Open Innov. Technol. Mark. Complex.*, 2018, doi: 10.3390/joitmc4030021.
- [3] M. Oliveira and D. Afonso, "Industry Focused in Data Collection: How Industry 4.0 is Handled by Big Data," in *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, 2019, pp. 12–18, doi: 10.1145/3352411.3352414.
- [4] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," *IEEE Internet Things J.*, 2016, doi: 10.1109/JIOT.2016.2579198.
- [5] H. N. Dai, H. Wang, G. Xu, J. Wan, and M. Imran, "Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies," *Enterp. Inf. Syst.*, 2019, doi: 10.1080/17517575.2019.1633689.
- [6] M. Blonda *et al.*, "Innovative Methodology for Detecting of Possible Harmful Compounds for Wastewater Treatment the MAUI Project," 2018, doi: 10.1109/METROI4.2018.8428315.
- [7] A. Gahr, P. Wazinski, and N. Andreas, "Water Management 4.0 in the Bitterfeld-Wolfen Chemical Park," *Chemie-Ingenieur-Technik*, 2019, doi: 10.1002/cite.201900011.
- [8] P. Nthutang and A. Telukdarie, "Integration of Small and Medium Enterprises for Industry 4.0 in the South African Water Services Sector: A Case Study for Johannesburg Water," 2019, doi: 10.1109/IEEM.2018.8607604.
- [9] M. Reis and G. Gins, "Industrial Process Monitoring in the Big Data/Industry 4.0 Era: from Detection, to Diagnosis, to Prognosis," *Processes*, vol. 5, no. 4, p. 35, Jun. 2017, doi: 10.3390/pr5030035.
- [10] L. Ribeiro, "Cyber-physical production systems' design challenges," 2017, doi: 10.1109/ISIE.2017.8001414.
- [11] E. A. Lee, "Cyber physical systems: Design challenges," 2008, doi: 10.1109/ISORC.2008.25.
- [12] Intel, "Connecting Connecting Legacy Legacy Devices Devices to the the Internet Internet of of Things Things ( IoT ) ( IoT )," 2014. [Online]. Available: <https://www.intel.com/content/dam/www/public/us/en/documents/solution-briefs/connecting-legacy-devices-brief.pdf>.
- [13] Y. Koren *et al.*, "Reconfigurable manufacturing systems," *CIRP Ann. - Manuf. Technol.*, 1999, doi: 10.1016/S0007-8506(07)63232-6.
- [14] Y. Koren and M. Shpitalni, "Design of reconfigurable manufacturing systems," *J. Manuf. Syst.*, 2010, doi: 10.1016/j.jmsy.2011.01.001.
- [15] M. G. Mehrabi, A. G. Ulsoy, and Y. Koren, "Reconfigurable manufacturing systems: key to future manufacturing," *J. Intell. Manuf.*, 2000, doi: 10.1023/A:1008930403506.
- [16] L. Restrepo, J. Aguilar, M. Toro, and E. Suescún, "A sustainable-development approach for self-adaptive cyber-physical system's life cycle: A systematic mapping study," *J. Syst. Softw.*, vol. 180, p. 111010, Oct. 2021, doi: 10.1016/J.JSS.2021.111010.
- [17] N. Papakostas, J. O'Connor, and G. Byrne, "Internet of things technologies in manufacturing: Application areas, challenges and outlook," 2017, doi: 10.1109/i-Society.2016.7854194.
- [18] E. Lee, Y.-D. Seo, and Y.-G. Kim, "Self-adaptive framework based on MAPE loop for internet of things," *Sensors (Switzerland)*, vol. 19, no. 13, 2019, doi: 10.3390/s19132996.
- [19] S. Zeadally, T. Sanislav, and G. D. Mois, "Self-Adaptation Techniques in Cyber-Physical Systems (CPSs)," *IEEE Access*, vol. 7, pp. 171126–171139, 2019, doi: 10.1109/ACCESS.2019.2956124.
- [20] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Commun. Mag.*, 2011, doi: 10.1109/MCOM.2011.6069707.
- [21] B. Bagheri, S. Yang, H. A. Kao, and J. Lee, "Cyber-physical systems architecture for self-aware machines in industry 4.0 environment," 2015, doi: 10.1016/j.ifacol.2015.06.318.
- [22] J. Lee, B. Bagheri, and H. A. Kao, "A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems," *Manuf. Lett.*, 2015, doi: 10.1016/j.mfglet.2014.12.001.
- [23] K. Schweichhart, "Reference Architectural Model Industrie 4.0 (RAMI 4.0) - An Introduction," *Platf. Ind. 4.0*, p. 21, 2016, [Online]. Available: <https://www.amnytt.no/getfile.php/3901260.2265.akzillql7uuipz/RAMI4+4.0.pdf>.
- [24] M. Paiva, A. Vasconcelos, and B. Fragoso, "Using enterprise architecture to model a reference architecture for industry 4.0," in *ICEIS 2020 - Proceedings of the 22nd International Conference on Enterprise Information Systems*, 2020, vol. 2, pp. 709–716.
- [25] Industrial Internet Consortium, "The Industrial Internet Reference Architecture v1.9," *Online*, 2019. <https://www.iiconsortium.org/pdf/IIRA-v1.9.pdf> (accessed Dec. 12, 2021).
- [26] E. Y. Nakagawa, P. O. Antonino, F. Schnicke, R. Capilla, T. Kuhn, and P. Liggesmeyer, "Industry 4.0 reference architectures: State of the art and future trends," *Comput. Ind. Eng.*, vol. 156, 2021, doi: 10.1016/j.cie.2021.107241.

Article

# A Cyber-Physical Data Collection System Integrating Remote Sensing and Wireless Sensor Networks for Coffee Leaf Rust Diagnosis

David Velásquez <sup>1,2,3,\*</sup> , Alejandro Sánchez <sup>1</sup>, Sebastián Sarmiento <sup>1</sup>, Camilo Velásquez <sup>1</sup>, Mauricio Toro <sup>1</sup> , Edwin Montoya <sup>1</sup> , Helmuth Trefftz <sup>1</sup> , Mikel Maiza <sup>2</sup>  and Basilio Sierra <sup>3</sup> 

- <sup>1</sup> RID on Information Technologies and Communications Research Group, Universidad EAFIT, Carrera 49 No. 7 Sur-50, Medellín 050022, Colombia; asanch41@eafit.edu.co (A.S.); ssarmien@eafit.edu.co (S.S.); cvelas31@eafit.edu.co (C.V.); mtorobe@eafit.edu.co (M.T.); emontoya@eafit.edu.co (E.M.); htrefftz@eafit.edu.co (H.T.)
  - <sup>2</sup> Department of Data Intelligence for Energy and Industrial Processes, Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), 20014 Donostia-San Sebastián, Spain; mmaiza@vicomtech.org
  - <sup>3</sup> Department of Computer Science and Artificial Intelligence, University of Basque Country, Manuel Lardizabal Ibilbidea, 1, 20018 Donostia-San Sebastián, Spain; b.sierra@ehu.es
- \* Correspondence: dvelas25@eafit.edu.co



**Citation:** Velásquez, D.; Sánchez, A.; Sarmiento, S.; Velásquez, C.; Toro, M.; Montoya, E.; Trefftz, H.; Maiza, M.; Sierra, B. A Cyber-Physical Data Collection System Integrating Remote Sensing and Wireless Sensor Networks for Coffee Leaf Rust Diagnosis. *Sensors* **2021**, *21*, 5474. <https://doi.org/10.3390/s21165474>

Academic Editor: Paolo Bellavista

Received: 13 July 2021

Accepted: 11 August 2021

Published: 13 August 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Coffee Leaf Rust (CLR) is a fungal epidemic disease that has been affecting coffee trees around the world since the 1980s. The early diagnosis of CLR would contribute strategically to minimize the impact on the crops and, therefore, protect the farmers' profitability. In this research, a cyber-physical data-collection system was developed, by integrating Remote Sensing and Wireless Sensor Networks, to gather data, during the development of the CLR, on a test bench coffee-crop. The system is capable of automatically collecting, structuring, and locally and remotely storing reliable multi-type data from different field sensors, Red-Green-Blue (RGB) and multi-spectral cameras (RE and RGN). In addition, a data-visualization dashboard was implemented to monitor the data-collection routines in real-time. The operation of the data collection system allowed to create a three-month size dataset that can be used to train CLR diagnosis machine learning models. This result validates that the designed system can collect, store, and transfer reliable data of a test bench coffee-crop towards CLR diagnosis.

**Keywords:** coffee leaf rust; cyber-physical system; internet of things; mechatronic design; technological integration; remote sensing; wireless sensor networks

## 1. Introduction

Coffee, for over 1000 years and even today, has been one of the most consumed drinks around the world with more than 400 billion cups per year [1]. Among more than 100 existing coffee species, only two are used for the drink preparation, namely *Coffea arabica* and *Coffea robusta*. The first one, which is used to obtain a more aromatic and softer beverage, is best valued by the market and represents over 75% of the world production. The drink resulting from processing the second one, which is considered to have a stronger and more bitter flavor, represents the remaining 25% [2]. Moreover, each species subdivides into coffee varieties, each of them having characteristics that allow the creation of distinct aromas and tastes.

As a case study, we consider the case of Colombia, which is the third major coffee producer of the world [3]. Colombia is located on the Bean Belt, a strip across the globe where all coffee plants are grown [4]. The national production is concentrated on *Coffea arabica*, due to the mountainous topography of the country, which offers a suited combination of altitude, temperature, and rainfall for this species. Particularly, the most cultivated

varieties of *Coffea arabica* in the country are Castillo, Colombia, Caturra, and Bourbon [5]. Depending on the selected variety and the post-harvesting process, the resulting product is offered in two different markets. One of them is the standard coffee market, which is guided by the international coffee price, and the other one is called the specialty coffee market, which has a premium above the standard price.

Regarding the phytosanitary problems on coffee crops, one of the main concerns is related to the presence of pests, such as the Coffee Borer Beetle, and diseases, such as the Coffee Brown Eye Spot and the Coffee Leaf Rust (CLR) [6]. For the diseases, the CLR is the most relevant one, in economic and pathological terms, at the national level. This disease presents a vertiginous expansion on the coffee plant and its surroundings, and it can cause massive defoliation on the whole crop [7]. As an example, in extreme cases, this disease has led to devastating losses in some Colombian regions reaching between 70% and 80% of the harvest [6].

It should be noted that the use of technology to support agriculture has made it possible to automate and optimize production. In this sense, sensors can be used both to monitor the machinery required for a plantation (e.g., performing predictive maintenance [8]) and to detect specific features of a crop and its ecosystem (e.g., non-invasive phenotyping in plant breeding [9]). Other applications of the use of sensors in agriculture may include precision irrigation, greenhouse instrumentation, and pest control [10].

It is noteworthy that, at the beginning of this research, the general objective was oriented to the early detection of the Coffee Brown Eye Spot disease through Remote Sensing (RS) with spectral reflectance data analysis. However, after carrying out the interviews with the Colombian coffee experts and producers, we realized that the mentioned disease was not as crucial or economically limiting as the CLR. The interviewees expressed that their main concern was the CLR and most of them even reported that they have been struggling with it over the last four years. Thus, and thanks to their recommendations, we decided not only to diagnose the CLR instead of the Coffee Brown Eye Spot disease, but also to integrate Wireless Sensor Networks (WSN). In that sense, the first step towards diagnosing the disease consisted of collecting reliable data regarding its development. Thereby, once the necessary data had been collected, it would be possible to create a diagnostic model based on such data. Therefore, this research presents the following two contributions: (i) The mechatronic design of a cyber-physical data collection system to collect and store data, integrating RS and WSN; (ii) a three-month dataset for CLR detection.

This paper is structured as follows: Section 2 explains some key concepts and describes related work by different authors, Section 3 presents the conceptual and detailed design of the data collection system, Section 4 shows the building and integration of the mechanical, electronic, and computing components and, finally, Section 5 states the conclusions, recommendations, and future work.

## 2. State of the Art

Different studies have been carried out involving technical methods and strategies for obtaining nutritional information of different types of crops [11], diagnosing diseases [12], and detecting pests [13]. Recently, an important concept has emerged called Precision Agriculture (PA). PA refers to an agricultural management concept that uses information and communications technology to observe, measure, and respond to specific crop variabilities. PA includes applying the correct treatment method at the right time according to the needs of the plants [14].

In PA, one of the current methods used to evaluate the features of different crops is called RS. RS relies on the interaction between materials and their electromagnetic radiation. It includes receiving radiation reflected from soil or plants to obtain valuable information, such as chlorophyll content, water stress, weed density, crop nutrients, and disease presence. These measurements can be made using airplanes, portable sensors, satellites, tractors, and drones [15].



Several authors [16–18] pointed out the importance of using high-quality portable devices to detect and control diseases in hard-to-reach sites. For example, Goel et al. [16] analyzed the detection of variations in the spectral response of corn (*Zea mays*) due to nitrogen application rate and weed control. For this reason, a hyperspectral sensor called Compact Airborne Spectrographic Imager is used to analyze the reflectance values of 72 bands in the range of 409 nm to 947 nm. These bands include visible light and external Near-Infrared (NIR) from the radiation spectrum. Their research demonstrated the potential of using hyperspectral sensors to detect weed infestation and nitrogen stress. Specifically, the most suitable wavelength bands for detection are found to be the wavelength regions around 498 nm and 671 nm, respectively.

In addition, a crop classification method employing the infrared and visible portions of the electromagnetic spectrum and low-cost cameras in a multi-rotor aircraft was proposed by Bolaños et al. [17]. This study is based on the identification of Normalized Difference Vegetation Index to assess health status and moisture content. Similarly, Chemura et al. [18] presented a method for predicting the presence of diseases and pest infestations early in coffee trees due to imperceptible water pressure. To this end, a handheld multi-spectral scanner with the visible and near-infrared regions is placed in an Unmanned Aerial Vehicle. Chemura et al. research is also related to irrigation planning based on the specific water needs of plants.

In addition to RS, based on smart farming techniques and the Internet of Things (IoT), which refers to the use of intelligently connected devices and systems leveraging data acquired by embedded sensors and actuators in machines and other physical objects [19], there is another popularly used method named WSN. WSN is responsible for real-time monitoring of different agricultural characteristics. It consists of multiple integrated, unattended devices called sensor nodes, which collect data at the site and wirelessly transmit it to a centralized processing station (called a base station). This station can store, process, and transmit data to the Internet, where a final user can analyze and transform it into relevant information and knowledge [20].

In this regard, Chaudary et al. highlighted the importance of WSN in the PA field by controlling and sensing the most relevant variables of a greenhouse using a microcontroller technology named Programmable System on Chip. This research examined the integration of wireless sensor nodes with high-bandwidth spectrum telecommunications technology, which proved helpful in determining the optimal irrigation strategy that meets crops' specific needs. Moreover, the study recommended using reliable low-current consumption hardware for WSN applications because it improves farmers' confidence to incorporate them into their crops [21]. Additionally, Piamonte et al. implemented a WSN prototype for monitoring an African Oil Palm disease called the Bud Rot. By using humidity, pH, light, and temperature sensors, their prototype measured climate change and soil factors to identify the presence of disease-causing fungi indirectly. This research concluded that the measurement results for the aforementioned non-biological factors had changed slightly, which, according to the researchers, indicates the possibility of detecting the Bud Rot [22].

The presented state of the art shows that RS and WSN are two widely used methods within PA due to their capacity to monitor different crop features and detect the presence of various anomalies.

### 3. System Design

Previous research was helpful to design a cyber-physical data collection system that could integrate both methods towards the CLR diagnosis. Applying the concepts and following the recommendations found in the state-of-the-art, it is possible to create a system capable of acquiring and remotely storing reliable data from diverse sources. The goal of such cyber-physical systems (CPS) is the characterization of a test bench coffee-crop regarding the changes induced by the disease at hand. The cyber-physical data collection system was designed following the Pahl and Beitz methodology [23]. The mechatronic design of the data collection system is presented in this section.

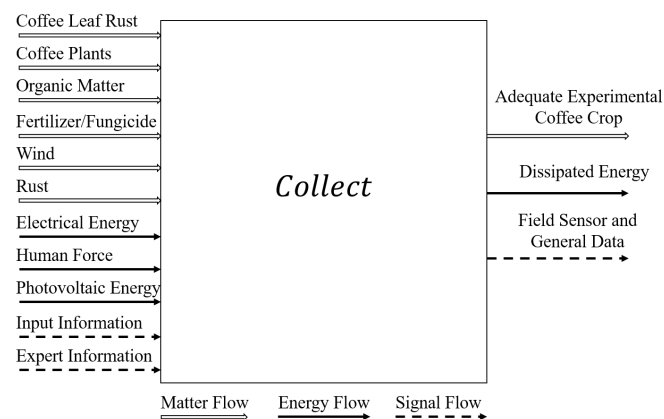
For the development of the system, requirements were elicited with the participation of Colombian Coffee Agricultures Association (CENICAFÉ) and University EAFIT. From EAFIT, the design team was composed by the Mechanical, Informatics and Electronic Engineers, as well as Biologists. The fulfillment of those requirements, which included, among others, building a test bench coffee-crop, emulating different agronomic conditions, and allowing the data acquisition, storage, and transfer, was the guideline for the design of the system. In that sense, this section describes, in a stepwise fashion, the use of the Pahl and Beitz methodology for achieving a data collection system that integrates RS and WSN towards the CLR diagnosis.

### 3.1. Main Requirements

First, all requirements were formalized, structured, and classified according to their characteristics and priority through the employment of the Product Design Specification [24]. The main requirement was measuring physicochemical features of the plants as well as capturing Red-Green-Blue (RGB) and multi-spectral images of the test bench coffee-crop for storing all this data locally and remotely. Other requirements were related to plants' separation and irrigation, coffee variety to be used, construction materials, database type, and communication protocol with the field sensors.

### 3.2. Black Box Definition

The following step is to design a black box [25], which represents the primary function of the system to be developed. This primary function is to collect a set of inputs, transform them, and produce a set of outputs. As shown in Figure 1, the inputs and outputs are divided into three major flows: namely matter, energy, and signal. Regarding the inputs, the matter flow was composed by CLR, coffee plants, organic matter, fertilizer/fungicide, and wind; the energy flow was divided into electrical energy, human force, and photovoltaic energy; and the signal flow consisted of input information and expert information. At the output, the adequate experimental coffee crop dissipated energy as well as field sensors and general data were obtained. These output data correspond to the main objective of this research, which is to create a system capable of collecting, storing, and transferring reliable data of a test bench coffee-crop towards the CLR diagnosis.



**Figure 1.** Black box representation of the cyber-physical data collection system.

### 3.3. Functional Structure

After defining the Black box, the functional structure [26] was specified, breaking down the presented inputs and outputs and establishing, with a detailed understanding, the sub-functions required, and the pathway created by these. As a way of example, one of these sub-functions consisted of merging human force with the coffee plants to arrange the

latter in the test bench coffee-crop, which impacts the posterior incorporation of the field sensors. The general pathway of the overall functional structure is described as follows.

The coffee plants were divided into four lots, where half of them were inoculated with *Hemileia vastatrix* [27], the fungus that causes CLR. For their agronomic management, fertilizer and fungicide were distributed and incorporated into all four lots. Then, each lot was isolated from the others to make them independent and, finally, the whole crop was integrated with the rain and wind emulation systems. Rainfall and wind speed were both monitored and regulated for the entire crop.

Furthermore, employing sensors in each lot, soil moisture/temperature, pH, illuminance, and environmental humidity/temperature were acquired. In addition, RGB and multi-spectral images were captured. To finish the data collection process, data were stored locally, pre-processed for cleaning purposes, and then sent to a remote server over the Internet. In addition, the collection process was monitored in real time through an IoT web platform.

#### 3.4. Morphological Matrix and Candidate Concepts

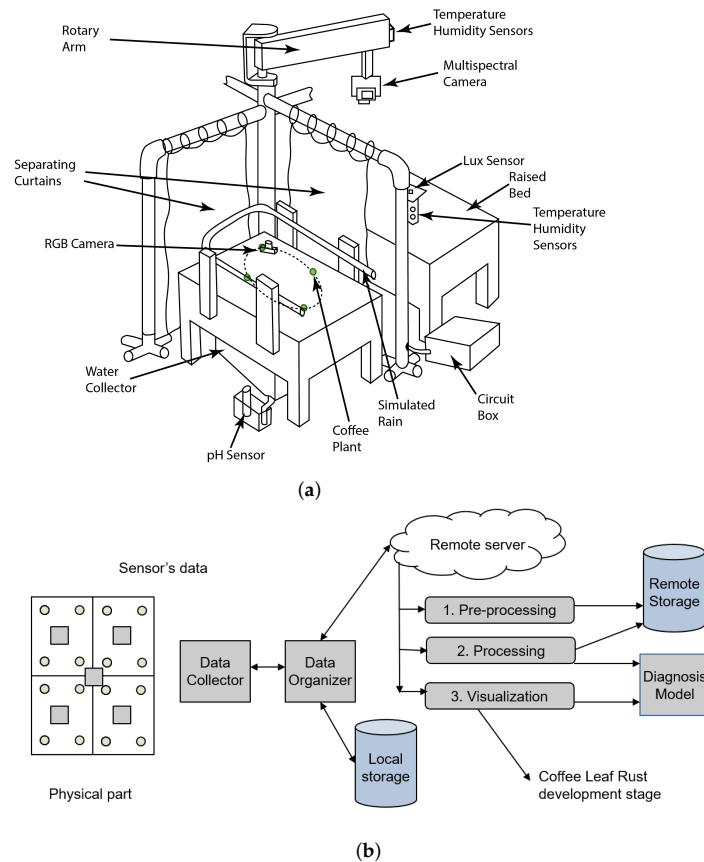
Once the main function and the corresponding sub-functions were specified, the morphological matrix [28] was developed. Such a matrix illustrates different solution proposals for the implementation of each of the sub-functions exposed in the functional structure. The output of the morphological matrix consisted of two candidate concepts, Concept 1 and Concept 2, each made up of a different combination of the solution proposals. The concepts indicated two possible ways of building the data collection system, and they were elaborated with the purpose of evaluating them under different aspects and deciding which was the most appropriate for the objective at hand. The most relevant features for Concept 1 and Concept 2 included: (i) holes in tubes or sprinklers to emulate rain, (ii) a stepper motor or a servomotor to position the multi-spectral cameras over the lots, (iii) using normal pressure from the aqueduct or a pump to transport the water for irrigation, and (iv) a rotary arm or a single rail to capture images from multispectral cameras, respectively (see Table S1 from supplementary material uploaded at MDPI platform and at provided link in Supplementary Materials section). The resulting candidate concepts were then evaluated by using a scoring system, which calculates a weighted average of a set of pre-selected evaluation criteria. These weights were established according to the previously defined PDS and the design team expertise. As a result, the final concept is selected. As shown in Table 1, Concept 1 resulted as the selected concept, with an approval of 78% against 74% of Concept 2. The cyber-physical data collection system was built based on the winning concept. CPS are a new class of engineered systems which offer close interaction between cyber and physical components [29]. It should also be noted that the chosen concept was slightly modified following some improvements proposed by CENICAFÉ and the design team to better fulfill the initial requirements.

#### 3.5. Final Concept

Figure 2a shows a sketch of the final concept for building the physical part of the system. This concept is composed of four raised wooden beds representing the lots and separated by four plastic curtains, a rotary arm holding the multi-spectral cameras, a rain system which irrigates the lots, and a circuit box with the necessary elements to interact with the electronic components. Figure 2b shows a sketch of the cybernetics part of the design, which includes a data collector for joining the data coming from the test bench coffee-crop and a data organizer, which structures and saves it on the local storage for its posterior transfer to a remote server located at EAFIT University.

**Table 1.** Concept Scoring. <sup>a</sup> Value scale (score between 0–4); 0 = Not satisfied, 1 = Acceptable, 2 = Sufficient, 3 = Good, 4 = Totally satisfied.

N <sup>o</sup>	Evaluation Criteria	Relevance (%)	Solutions <sup>a</sup>	
			Concept 1	Concept 2
1	Functionality	11	4	3
2	Simplicity	5	3	4
3	Fulfilment of requirements	10	3	2
4	Robustness	3	4	3
5	Fabrication	7	3	3
6	Assembly	6	3	2
7	Reliability	9	3	3
8	Low cost	7	3	3
9	Expert criteria	6	3	3
10	Crop management	7	3	3
11	Maintainability	3	2	3
12	Performance	8	2	3
13	Usability	5	3	3
14	Testability	3	3	2
15	Availability	10	4	4
	Weighted average		3.13	2.96
	Total score	100	78%	74%



**Figure 2.** Final concept sketch for the data collection system: (a) of the physical part; (b) of the cybernetic part.

### 3.6. Mechanical Design

The mechanical design considered four identical crop lots, each one housing four coffee plants under different development stages of the disease. Each lot had specific structures designated to place the sensors for pH, illuminance, soil moisture, soil temperature, environmental humidity, and environmental temperature. In addition to the sensors, a rotary platform, holding a rack-pinion mechanism and containing a slider extension, three micro-switches, a mini-DC motor, and a digital servo, was also designed for each lot aiming at driving an RGB camera close to each plant for capturing images from the bottom of the leaves. In addition, each lot had a filtering point, which directed the residual water into a container where the pH-meter was placed.

Furthermore, a rotary arm was designed to place two multi-spectral cameras and to be able to move them above the four crop lots for capturing images from the upper side of the leaves. One of the cameras had an RGN filter (Red-850 nm, Green-660 nm, Near Infrared-550 nm), whereas the other one had a RE filter (Red Edge-735 nm). Both filters were suitable for assessing the presence of plagues and diseases (in particular the CLR) in crops [30,31]. Moreover, since the coffee plants needed a suitable environment to grow, a rain system was also designed for irrigation purposes. This system was controlled through an open/close command that could change the state of a corresponding solenoid valve according to a pre-defined rain schedule.

### 3.7. Electronic Design

To collect data from each crop lot using field sensors, an electronic design of the system was required. Sensors, actuators, interfaces, power supplies, two (2) Arduino Mega microcontrollers, one (1) Raspberry Pi microcomputer, and an electrical cabinet composed this electronic design. These electronic components were connected and integrated to support the cybernetic part of the data collection system. In what follows, each component is described.

#### 3.7.1. Arduino Mega Microcontrollers

One of the Arduino Mega microcontrollers (named Arduino 1) collected lot data. Thus, four pH-meters, four illuminance sensors, four soil moisture sensors, four soil temperature sensors, and one environmental humidity/temperature sensor were connected to it. In addition, the Arduino controlled the movements of the central rotary platform. For its part, the other Arduino Mega (named Arduino 2) was considered for the general data collection, having the tasks of activating/deactivating the rain system, communicating with the flow and wind sensors, as well as moving the rotatory arm over the lots. Both Arduino Mega microcontrollers were communicated with the Raspberry Pi via USB.

#### 3.7.2. Raspberry Pi

The Raspberry Pi was responsible for orchestrating the sequence of steps during each data collection routine, storing the gathered data locally, and transferring it to a remote server over the Internet. For that purpose, in addition to being communicated with both Arduino Mega microcontrollers, four RGB and two multi-spectral cameras were connected to it via USB. Thereby, the Raspberry Pi was able to trigger the different electronic devices and obtain the results. Finally, to send data to the remote server, an outdoor 4G LTE router was used to facilitate remote connectivity from the Data Collection System location.

#### 3.7.3. Electrical Cabinet

The design of an electrical cabinet was required for the distribution and organization of all electronic components. This cabinet had an IP5 minimum environmental protection due to the system exposure to the greenhouse's harsh conditions. Furthermore, a current security breaker was also included to protect the components from a peak current over 10 A, considering that the total consumption of the data collection system was about 7 A. Additionally, protection fuses were proposed for each power supply and actuator to miti-

gate damages, and one 9 V/1 A AC/DC adapter was considered for each microcontroller to avoid problems related to low current values.

The Raspberry Pi and the Arduino Mega microcontrollers were connected through a Master–Worker network architecture. The connections were implemented by a serial interface.

### 3.8. Software Design

The software design was essential to detail how the physical components communicated with the cybernetic part of the system to collect and transfer the data properly. In that sense, it is essential to explain the principal functions, commands, components, architectures, content specifications and platforms, which were thought as necessary for managing the incoming and outgoing data flow.

#### 3.8.1. Data Acquisition, Conditioning, and Storage Routines

Several functions regarding the acquisition, conditioning, and storage of the readings of the electronic devices connected to Arduino 1 and Arduino 2 were defined. Arduino 1 functions were in charge of collecting, conditioning, and storing the data proceeding from the pH, soil temperature/moisture, illuminance and environmental temperature/humidity sensors, and controlling the servomotor angle and arm's extension to position the RGB camera of each lot. Arduino 2 functions included collecting, conditioning, and storing readings from the flow rate and wind speed sensors and controlling the direction and destination of the global rotary arm that holds the multispectral cameras. The programs of both Arduino Mega microcontrollers were designed to respond to specific commands sent by the Raspberry Pi.

#### 3.8.2. Main Orchestration Program

A main, global program run by a Raspberry Pi in charge of orchestrating every step of the data collection routines and automatically executing such a process seven times a day was defined. Such an orchestration program was also in charge of activating/deactivating the rain system according to a pre-set schedule. This pre-set schedule included the raining days of year 2018. In addition, the program implemented by the Raspberry Pi was in charge of receiving and organizing the collected data from the Arduino Mega microcontrollers, triggering the RGB and multi-spectral cameras, storing everything locally in a structured way and transferring it to a remote server, named Academic Data Center (ADC), over the Internet via Secure File Transfer Protocol (SFTP). Finally, this program reported the current state to Thingworx (IoT platform, see subsection (iii)) during each routine. The software technologies selected for the implementation of this program were Python 3.5.3, OpenCV 3.4.1 [32], RPi.GPIO 0.6.3 [33] and MongoDB [34].

#### 3.8.3. Thingworx

*Thingworx* is a complete software platform designed for the Industrial IoT [35]. It was used to develop a dashboard, to remotely monitor the field and general data in real-time. This software platform was installed in the ADC. The ADC is the remote server used for remotely storing the collected data from the CPS to replicate the Raspberry Pi's local storage. This server is hosted by the Computer-Science Department of *EAFIT University* and is composed of 72 Cores, 512 GB RAM, 4 TB Storage, 2 GPU Tesla K80, and an Ubuntu 18.04 Operating System. It can be accessed over the Internet through a *Virtual Private Network (VPN)* connection.

The use of the Pahl and Beitz methodology allowed for evolving from the elicitation of the initial requirements towards the achievement of the final and detailed design of a data collection system that integrates RS and WSN. The final 3D Computer-aided design (CAD) of the data collection system's physical part is shown in Figure 3.

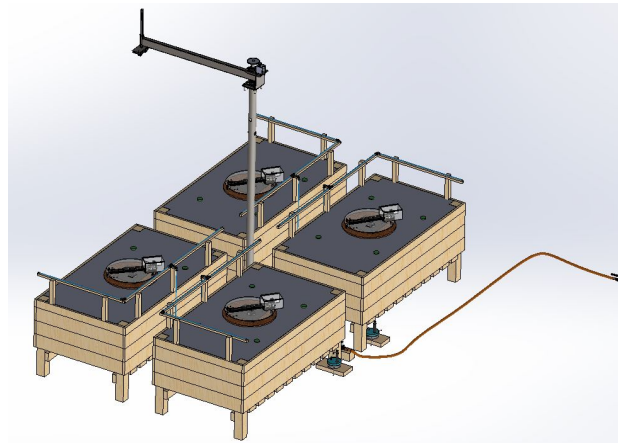


Figure 3. Final 3D CAD of the data collection system’s physical part.

Figure 4 shows the final design of the data collection system’s cybernetic part, which explains the pipeline for the execution of the data collection system.

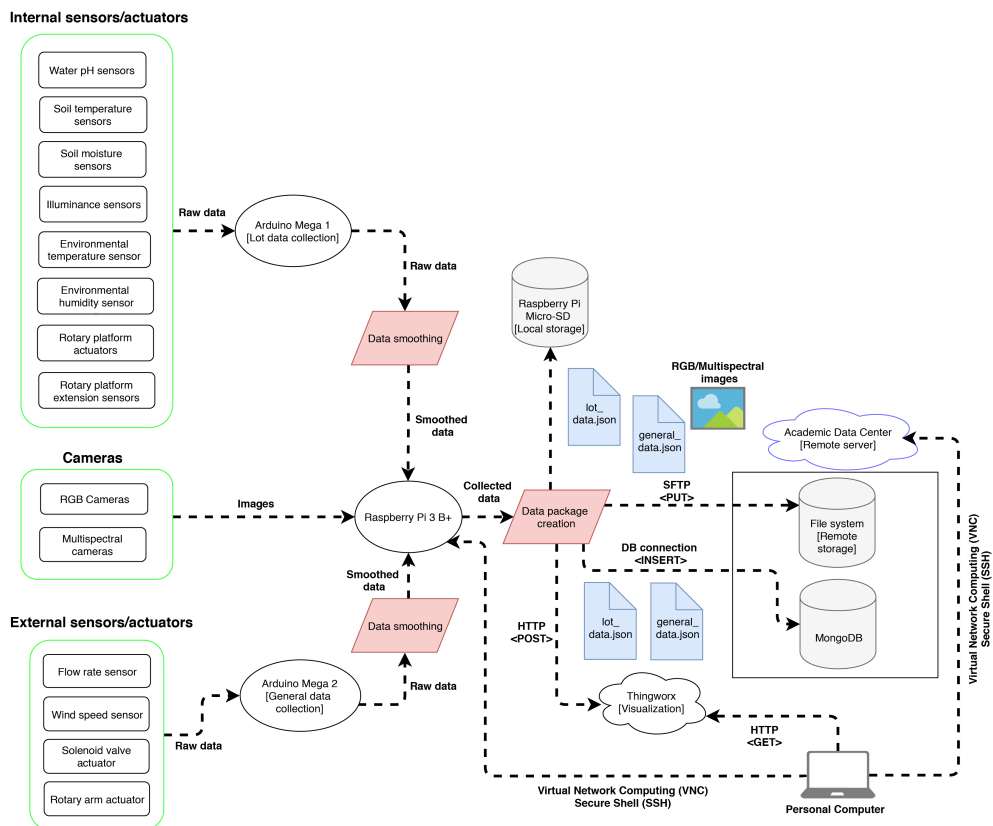


Figure 4. Final design of the data collection system’s cybernetic part.

## 4. Results

The application of the Pahl and Beitz design methodology results in the target system solution. Precisely, the result of this research work corresponds to the solution obtained by applying the Pahl and Beitz methodology, i.e., the target system, i.e., a cyber-physical data acquisition system capable of obtaining a dataset suitable for use in the early detection of CLR. Therefore, the following describes the results obtained, i.e., the system built, which was first represented by a 3D-CAD model.

### 4.1. Mechanical Components

These included the installation of a set of curtains, which were planned to work as a separation between the four crop lots. In addition, the assembly of the rotary arm structure was carried out. Finally, a separate structure for holding the multi-spectral cameras container was fixed.

Additionally, due to the emulated rain conditions that the prototype would be subjected to, immunized wood was chosen for the construction of the coffee crop lots, since it is resistant to moisture. At the bottom of each lot, a mesh in conjunction with a plastic tarp was installed to contain the soil. Furthermore, a slope was built within each lot using soil and impermeable plastic with the purpose of driving the residual water into the pH measurement system.

Regarding the rain system, all accessories related to the main pipeline for the incoming water source were installed over immunized wooden blocks. In addition, the whole wooden base was buried to keep the structure fixed. In addition, five supports were also installed on two adjacent sides of each lot for mounting a wooden L-structure over them. That structure served as a base for three additional hoses with small perforations, which would distribute the water within the lot producing the rain effect.

Succeeding the plants' irrigation, the slope, which was built within each lot using soil and impermeable plastic, was helpful to drive the residual water into the filtering point located in one corner of each coffee crop lot. There, a hose was connected to lead the water to the container for the pH measurement.

Concerning the assembly of the rotary platform, several acrylic pieces were cut, including the rotary platform, rack, pinion, support for the rack, base, and protection for the RGB camera, supports for the mini-DC motor and digital servo and mechanical end stop for the rotary platform's extension. To achieve the movement of the latter, a drawer slide was installed below this extension, and, therefore, the camera displacement towards the coffee plants could be achieved. Furthermore, before placing the rotary platforms in the center of each lot, steel plates were assembled to the supports of the digital servos and four levelers were screwed to the corners of each plate to put them underground for fixing the whole structure. Lastly, each platform was placed on top of a wooden base to mitigate the terrain's instability.

### 4.2. Electronic Components and Their Integration

Following the construction of the mechanical components, the electronic integration was executed to complete the system's physical part. For that purpose, each sensor and actuator were tested, calibrated, and connected to the corresponding micro-controller. In addition, individual connectors with a thick silicon protective layer were added to each of them to keep their metal terminals safe from the harsh conditions, which would include high temperatures, soil, and water from the rain system on a regular basis.

Similarly, the sensors' calibration process was carried out to ensure the correct measurement and reliability of the data to be collected. Some sensors were already pre-calibrated at the factory (e.g., ambient temperature and humidity sensors) while others required calibration. For example, one of these sensors was the pH sensor, for which two buffer solutions with exact pH values were used. The sensor was adjusted to the same measured pH value as the buffer solution using the included potentiometers in the sensor's interface.



The calibration of the RGB cameras was carried out by manually adjusting the lens's focus with respect to the leaves of the coffee plant to obtain a correct image sharpness.

After verifying the proper functioning and calibration of the sensors and actuators in the laboratory, along with the micro-controllers, they were merged into the corresponding mechanical structures to form the integrated components, which were required for the general operation of the cyber-physical data collection system. In addition, to protect, contain, and distribute the completely intermediate electronic devices (Raspberry Pi, Arduino Mega microcontrollers, USB-hubs, power supplies, and interfaces), an electrical hermetic cabinet was employed. To this end, the cabinet was subdivided into different sections, and it was also tailored for offering protection against dust, water, and a possible peak current. After installing it, the sensors, actuators, and cameras were connected to their corresponding place inside the electrical cabinet, guiding their cables through impermeable PVC pipes coming from the lots.

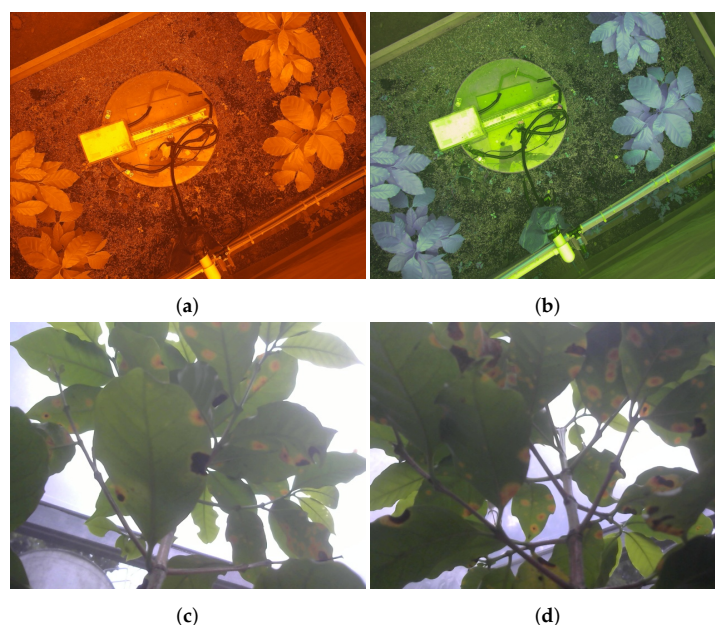
#### 4.3. Software Components

Once all electronic components were duly tested and the respective drawbacks were successfully solved, it was possible to send commands to the Arduino Mega microcontrollers and trigger the cameras from the Raspberry Pi with the objective of verifying the communication, checking that the desired actions were executed and validating the results. Thereby, the integration of the mechanical and electronic components, along with the ability to control the system from the Raspberry Pi, was successfully proven.

For its part, the cybernetic part of the data collection system began with the implementation of the communication between Raspberry Pi and Arduino microcontrollers. It was achieved through the development of an Arduino communicator component, which established two separate serial connections and grouped the responses of the micro-controllers into single programming objects that could be subsequently structured and stored. Moreover, a data sub-directory creator component was developed for creating a sub-directory within the main data directory using the timestamp at which each data collection routine began. This component was also in charge of generating the proper internal structure for the files, which consisted of one sub-directory for each lot and another one for the general data. With respect to the files, different software components were also implemented to capture the RGB and multi-spectral images of the plants and write corresponding JSON files with the collected data and the paths to those images. Figure 5 shows an example of some of the generated files after concluding a routine.

In addition, a data visualization dashboard was developed using Thingworx according to the presented design for monitoring the current state during each routine. Thingworx's appearance was preserved, and it was accessible through a Uniform Resource Locator (URL) with user-password authentication. Every used widget for creating the dashboard had a unique identifier so that it was possible to target each of them separately for updating their values.

After a routine finished, another software component, named data uploader, was responsible for transferring the collected data to the ADC over the Internet using a Wi-Fi connection. Consequently, the procedure to verify that the result was satisfactory consisted of checking whether the files uploaded to the ADC were identical to those stored locally in the Raspberry Pi.



**Figure 5.** Example of generated files after data collection routine: (a) RGN image from lot 1; (b) RE image from lot 1; (c) RGB image from plant 3, lot 3; (d) RGB image from plant 4, lot 3.

#### 4.4. Global Integrated System

Finally, the integration of the mechanical, electronic, and software parts led to the construction of a complete functional cyber-physical data collection system. The approximate total cost of the system implemented with all its components was 4863 USD. The total for the electronic and computer components were 3535 USD and for the mechanics 1328 USD. The most expensive components, in general, were the multispectral cameras with a cost of 1318 USD. It is estimated that, by implementing this system, followed by integrating an ML-based early CLR detection predictive model, the current harvesting losses due to this disease (e.g., 70% to 80% in Colombia) could be halved. Regarding maintenance costs, it is similarly estimated that these will be below 1% of net gains, which, considering estimated savings, is entirely affordable. Finally, the system's lifetime is estimated to be five to ten years, which is usual in this kind of equipment and is within the standards of amortization periods. Nevertheless, these figures should be validated for a full-scale implementation of the system.

After completing the integration and construction of the data acquisition system mentioned above, a final test and calibration of each system component's operation was performed, which is essential to ensure the system's operation's reliability. For example, one part of this process included the precise adjustment of the robotic arm position with respect to each plants' lot for taking the multispectral photos, where each position was stored in the program to perform the data collection routine. Having performed the final system's calibration, a data collection routine was executed for three months. The Data Collection System recorded crop's cameras and sensors information from each lot seven times per day at different moments (with and without sunlight). It must be noted that, although the data storage occurred seven times per day, the system was acquiring and monitoring (Thingworx) the data in real-time, with a sampling period of 3 s. In addition to the data collected by the system, a biologist team evaluated and labeled daily in a separate file the current development stage of the CLR of each data collection system lot. The output of this routine generated a dataset comprising 603 RGN files (~153 MB), 641 RE files (~177 MB), 730 RGB files (~196 MB), and 672 sensor data (JSON) files (~1.12 MB), which

were ready to be used for diagnosing the CLR development stage by training a Machine Learning model.

The operation of the current data collection system allowed the creation of a three-month size dataset. This dataset was used to train a deep learning model based on an ensemble algorithm integrating three convolutional neural networks and a multi-layer perceptron fed by RGB, RGN, and RE images; and Wireless Sensor Network data, correspondingly. This model was used to classify the early stage of CLR of a coffee crop (from 0 to 4), obtaining an F1-score of 0.775 [36].

## 5. Conclusions

This paper presents the mechatronic design of a cyber-physical data collection system, which integrates RS and WSN on a test bench coffee-crop. It is capable of automatically collecting, structuring, and locally and remotely storing reliable multi-type data from different field sensors (pH, soil moisture/temperature, illuminance, and environmental humidity/temperature), RGB and multi-spectral cameras. In addition, a data visualization dashboard was implemented to monitor the data collection routines in real time. This result represents a first step towards the CLR diagnosis on the *Caturra* variety.

The correct operation of the data collection system allowed for creating a three-month size dataset, which contains sensors and camera data required for creating a CLR development stage model. This result validates that the designed system can collect, store, and transfer reliable data of a test bench coffee-crop towards the CLR diagnosis.

For future work, this data collection system may be useful for measuring and recording different characteristics from other types of crops. In addition, and regarding the CLR, the data acquired through this system can be exploited for analyzing how the crop responds (in physicochemical and visual terms) according to the presence of the disease. It could be considered, for instance, to implement Artificial Intelligence techniques, such as Computer Vision and Deep Learning, to create a model based on the collected data for effectively diagnosing the CLR.

The current development is intended to be used as a test laboratory for plant experiments, which means that the obtained results are limited to a sample of a real crop plantation. As future work, a scalability, cost, and power consumption analysis could be carried out to turn the test laboratory into a full-scale mobile system. No relevant limitations are identified; however, employing drones and land robots are considered a technological requirement. Regarding drones, multispectral cameras (RGN and RE), which show the CLR in a distinct color from a top view of the crop, should be used. Concerning land robots, to effectively detect the CLR, they should be equipped with RGB cameras to monitor CLR's yellow spots under the coffee leaves and land sensors (e.g., pH, temperature, humidity, soil moisture, and luminance). This research work will be helpful to size the optimal number and type of sensors required by such a full-scale implementation.

**Supplementary Materials:** The following is available online at <https://www.mdpi.com/article/10.3390/s21165474/s1>, Table S1: Morphological Matrix.

**Author Contributions:** Conceptualization and investigation, A.S., S.S., C.V. and D.V.; supervision, D.V., M.T., E.M. and H.T.; validation and methodology, M.M. and B.S. All authors contributed to the writing and reviewing of the present manuscript. All authors read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the University EAFIT under Grant 828-000010 and the Colombian Science and Technology Department (Colciencias) under Grant "Jóvenes Investigadores e Innovadores por la Paz 2017".

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** This paper has a three-month dataset generated by the CPS Data Collection System, available at <https://iee-dataport.org/documents/coffee-leaf-rust-dataset> (accessed on 8 July 2021).

**Acknowledgments:** The authors would like to thank University EAFIT for providing the funds for the present research. Also we would like to thank Vicomtech foundation for supporting article processing charges.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

ADC	Academic Data Center
CLR	Coffee Leaf Rust
CPS	Cyber-Physical Systems
NIR	Near Infrared
IoT	Internet of Things
PA	Precision Agriculture
RE	Red Edge
RGB	Red Green Blue
RS	Remote Sensing
SFTP	Secure File Transfer Protocol
URL	Uniform Resource Locator
VPN	Virtual Private Network
WSN	Wireless Sensor Networks







## References

- Mussatto, S.I.; Machado, E.M.S.; Martins, S.; Teixeira, J.A. Production, composition, and application of coffee and its industrial residues. *Food Bioprocess Technol.* **2011**, *4*, 661. [CrossRef]
- Etienne, H. Somatic embryogenesis protocol: Coffee (*Coffea arabica* L. and *C. canephora* P.). In *Protocol for Somatic Embryogenesis in Woody Plants*; Springer: Dordrecht, The Netherlands, 2005; pp. 167–179.
- Coffee Total Production*; Technical Report; International Coffee Organization: London, UK, 2019.
- The Influence of Coffee around the World*; National Coffee Association: New York, NY, USA, 2015.
- Arcila, J.; Farfan, F.F.; Moreno, A.M.; Salazar, L.F.; Hincapié, E. *Sistemas de Producción de Café en Colombia*; Cenicafé: Chinchiná, Colombia, 2007.
- Rivillas, C.A.; Serna, C.A.; Cristancho, M.A.; Gaitan, A.L. *La Roya del Cafeto en Colombia: Impacto Manejo y Costos del Control*; Technical Report; Cenicafé: Chinchiná, Colombia, 2011.
- Nutman, F.J.; Roberts, F.M. Coffee leaf rust. *Pans Pest Artic. News Summ.* **1970**, *16*, 606–624. [CrossRef]
- Pech, M.; Vrchota, J.; Bednář, J. Predictive Maintenance and Intelligent Sensors in Smart Factory: Review. *Sensors* **2021**, *21*, 1470. [CrossRef] [PubMed]
- Busemeyer, L.; Mentrup, D.; Möller, K.; Wunder, E.; Alheit, K.; Hahn, V.; Maurer, H.P.; Reif, J.C.; Würschum, T.; Müller, J.; et al. BreedVision—A Multi-Sensor Platform for Non-Destructive Field-Based Phenotyping in Plant Breeding. *Sensors* **2013**, *13*, 2830–2847. [CrossRef]
- Ruiz-Garcia, L.; Lunadei, L.; Barreiro, P.; Robla, I. A Review of Wireless Sensor Technologies and Applications in Agriculture and Food Industry: State of the Art and Current Trends. *Sensors* **2009**, *9*, 4728–4750. [CrossRef]
- Mirik, M.; Norland, J.E.; Crabtree, R.L.; Biondini, M.E. Hyperspectral one-meter-resolution remote sensing in Yellowstone National Park, Wyoming: I. Forage nutritional values. *Rangel. Ecol. Manag.* **2005**, *58*, 452–458. [CrossRef]
- Lobitz, B.; Beck, L.; Huq, A.; Wood, B.; Fuchs, G.; Faruque, A.S.G.; Colwell, R. Climate and infectious disease: Use of remote sensing for detection of *Vibrio cholerae* by indirect measurement. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 1438–1443. [CrossRef]
- Su, N.Y. Remote Monitoring System for Detecting Termites. U.S. Patent 6,052,066, 18 April 2000.
- JRC of the European Commission. *Precision Agriculture: An Opportunity for Eu Farmers-Potential Support With the Cap 2014–2020*. European Union: Brussels, Belgium, 2014; p. 56. [CrossRef]
- Mulla, D.J. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosyst. Eng.* **2013**, *114*, 358–371. [CrossRef]
- Goel, P.K.; Prasher, S.O.; Landry, J.A.; Patel, R.M.; Bonnell, R.B.; Viau, A.A.; Miller, J.R. Potential of airborne hyperspectral remote sensing to detect nitrogen deficiency and weed infestation in corn. *Comput. Electron. Agric.* **2003**, *38*, 99–124. [CrossRef]

17. Bolaños, J.A.; Campo, L.; Corrales, J.C. Characterization in the Visible and Infrared Spectrum of Agricultural Crops from a Multirotor Air Vehicle. In Proceedings of the International Conference of ICT for Adapting Agriculture to Climate Change, Popayán, Colombia, 22–24 November 2017; Springer: Cham, Switzerland, 2017; pp. 29–43.
18. Chemura, A.; Mutanga, O.; Dube, T. Remote sensing leaf water stress in coffee (*Coffea arabica*) using secondary effects of water absorption and random forests. *Phys. Chem. Earth Parts A/B/C* **2017**, *100*, 317–324. [[CrossRef](#)]
19. GSMA Association. Understanding the Internet of Things (IoT). *arXiv* **2014**, arXiv:1011.1669v3.
20. Azfar, S.; Nadeem, A.; Shaikh, A.B. Pest Detection and Control Techniques Using Wireless Sensor Network: A Review. *J. Entomol. Zool. Stud.* **2015**, *3*, 92–99.
21. Chaudhary, D.D.; Nayse, S.P.; Waghmare, L.M. Application of wireless sensor networks for greenhouse parameter control in precision agriculture. *Int. J. Wirel. Mob. Netw. (IJWMN)* **2011**, *3*, 140–149. [[CrossRef](#)]
22. Piamonte, M.; Huerta, M.; Clotet, R.; Padilla, J.; Vargas, T.; Rivas, D. WSN Prototype for African Oil Palm Bud Rot Monitoring. In Proceedings of the International Conference of ICT for Adapting Agriculture to Climate Change, Popayán, Colombia, 22–24 November 2017; Springer: Cham, Switzerland, 2017; pp. 170–181.
23. Pahl, G.; Wallace, K.; Blessing, L.T.M.; Beitz, W.; Bauert, F. *Engineering Design: A Systematic Approach*; Springer: London, UK, 2013.
24. Ma, X.J.; Ding, G.F.; Qin, S.F.; Li, R.; Yan, K.Y.; Xiao, S.N.; Yang, G.W. Transforming Multidisciplinary Customer Requirements to Product Design Specifications. *Chin. J. Mech. Eng.* **2017**, *30*, 1069–1080. [[CrossRef](#)]
25. Bunge, M. A General Black Box Theory. *Philos. Sci.* **1963**, *30*, 346–358. [[CrossRef](#)]
26. Liu, A.; Lu, S. Functional design framework for innovative design thinking in product development. *CIRP J. Manuf. Sci. Technol.* **2020**, *30*, 105–117. [[CrossRef](#)]
27. Avelino, J.; Muller, R.; Eskes, A.; Santacreo, R.; Holguin, F. La roya anaranjada del cafeto: Mito y realidad. In *Desafíos de la Caficultura en Centroamérica*; IICA: San José, Costa Rica, 1999; pp. 194–241.
28. Kang, Y.; Tang, D. Matrix-based computational conceptual design with ant colony optimisation. *J. Eng. Des.* **2013**, *24*, 429–452. [[CrossRef](#)]
29. Khaitan, S.K.; McCalley, J.D. Design techniques and applications of cyberphysical systems: A survey. *IEEE Syst. J.* **2015**, *9*, 350–365. [[CrossRef](#)]
30. Thenkabail, P.S.; Lyon, J.G.; Huete, A. Hyperspectral remote sensing of vegetation and agricultural crops: Knowledge gain and knowledge gap after 40 years of research. In *Hyperspectral Remote Sensing of Vegetation*; CRC Press: Boca Raton, FL, USA, 2016; pp. 698–763.
31. Chemura, A.; Mutanga, O.; Sibanda, M.; Chidoko, P. Machine learning prediction of coffee rust severity on leaves using spectroradiometer data. *Trop. Plant Pathol.* **2018**, *43*, 117–127. [[CrossRef](#)]
32. Pulli, K.; Baksheev, A.; Korniyakov, K.; Eruhimov, V. Real-time computer vision with OpenCV. *Commun. ACM* **2012**, *55*, 61–69. [[CrossRef](#)]
33. Chaczko, Z.; Braun, R. Learning data engineering: Creating IoT apps using the node-RED and the RPI technologies. In Proceedings of the 2017 16th International Conference on Information Technology Based Higher Education and Training (ITHET), Ohrid, Macedonia, 10–12 July 2017; pp. 1–8.
34. Alvermann, M. Introduction to MongoDB. 2016. p. 9. Available online: <https://www.mongodb.com/citedon> (accessed on 18 July 2020).
35. Mineraud, J.; Mazhelis, O.; Su, X.; Tarkoma, S. Contemporary internet of things platforms. *arXiv* **2015**, arXiv:1501.07438
36. Velásquez, D.; Sánchez, A.; Sarmiento, S.; Toro, M.; Maiza, M.; Sierra, B. A Method for Detecting Coffee Leaf Rust through Wireless Sensor Networks, Remote Sensing, and Deep Learning: Case Study of the Caturra Variety in Colombia. *Appl. Sci.* **2020**, *10*, 697. [[CrossRef](#)]

Article

# A Method for Detecting Coffee Leaf Rust through Wireless Sensor Networks, Remote Sensing, and Deep Learning: Case Study of the Caturra Variety in Colombia

David Velásquez <sup>1,2,3,\*</sup>, Alejandro Sánchez <sup>1</sup>, Sebastian Sarmiento <sup>1</sup>, Mauricio Toro <sup>1</sup>,  
Mikel Maiza <sup>2</sup> and Basilio Sierra <sup>3</sup>

<sup>1</sup> I+D+i on Information Technologies and Communications Research Group, Universidad EAFIT, Carrera 49 No. 7 Sur - 50, Medellín 050022, Colombia; asanch41@eafit.edu.co (A.S.); ssarmien@eafit.edu.co (S.S.); mtorobe@eafit.edu.co (M.T.)

<sup>2</sup> Department of Data Intelligence for Energy and Industrial Processes, Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), 20014 Donostia-San Sebastián, Spain; mmaiza@vicomtech.org

<sup>3</sup> Department of Computer Science and Artificial Intelligence, University of Basque Country, Manuel Lardizabal Ibilbidea, 1, 20018 Donostia/San Sebastián, Spain; b.sierra@ehu.eus

\* Correspondence: dvelas25@eafit.edu.co

Received: 20 December 2019; Accepted: 15 January 2020; Published: 19 January 2020



**Abstract:** Agricultural activity has always been threatened by the presence of pests and diseases that prevent the proper development of crops and negatively affect the economy of farmers. One of these pests is Coffee Leaf Rust (CLR), which is a fungal epidemic disease that affects coffee trees and causes massive defoliation. As an example, this disease has been affecting coffee trees in Colombia (the third largest producer of coffee worldwide) since the 1980s, leading to devastating losses between 70% and 80% of the harvest. Failure to detect pathogens at an early stage can result in infestations that cause massive destruction of plantations and significantly damage the commercial value of the products. The most common way to detect this disease is by walking through the crop and performing a human visual inspection. As a result of this problem, different research studies have proven that technological methods can help to identify these pathogens. Our contribution is an experiment that includes a CLR development stage diagnostic model in the *Coffea arabica*, Caturra variety, scale crop through the technological integration of remote sensing (through drone capable multispectral cameras), wireless sensor networks (multisensor approach), and Deep Learning (DL) techniques. Our diagnostic model achieved an  $F_1$ -score of 0.775. The analysis of the results revealed a  $p$ -value of 0.231, which indicated that the difference between the disease diagnosis made employing a visual inspection and through the proposed technological integration was not statistically significant. The above shows that both methods were significantly similar to diagnose the disease.

**Keywords:** coffee leaf rust; machine learning; deep learning; remote sensing; Fourth Industrial Revolution; Agriculture 4.0

## 1. Introduction

The food and beverage industry is characterized by a relatively small number of multinational companies that link small producers around the world with consumers. A development analysis conducted by the World Economic Forum and Accenture, in 2018 [1], focused, predominantly, on upstream value chain segments due to the low tech nature of food and beverage processing and production and the substantial potential for improving efficiency in agrifood activities.

According to the Organisation for Economic Co-operation and Development (OECD), the food and beverage industry is classified as a low tech industry, so it can add innovation without significant social disadvantages [2]. According to the OECD, each opportunity presented by the Fourth Industrial Revolution must be used to realize a global food production system that can address challenges with limited environmental impact while taking advantage of opportunities for growth, innovation, and development [2].

The developments of the Fourth Industrial Revolution will change production systems in the food and beverage industry through innovation in digital, physical, and biological technologies [1]; for instance, vertical agriculture, advanced wastewater treatment, advanced packaging, precision agriculture [3], advanced organic agriculture, supply chain traceability [4], genome editing, cell and tissue engineering, automated agriculture [5], remote sensing [6], 3D food printing, and Agriculture 4.0.

The three main developments with the most significant growth potential for value creation in the food and beverage industry are: precision agriculture, advanced organic agriculture, and genome publishing [1]. In particular, precision agriculture integrates data analysis processes with crop science and technologies such as GPS, soil sensors, meteorological data, and the Internet of Things (IoT) for decisions related to fertilizer, irrigation, harvest time, and seed spacing, among others. Precision agriculture is applicable to the entire agricultural production system and drives substantial yield increases while optimizing for resource use [1]. The goal of precision agriculture is to enable scientific decisions in agriculture to improve value creation.

One industry in which precision agriculture can improve value creation is the coffee industry; in particular, the specialty market. Coffee is one of the world's most popular drinks and merchantable commodities. Every year, over 500,000 million cups are consumed, and over 158 million bags of 60 kg are produced. Coffee is grown in around 70 countries around the world in a region known as the Bean Belt. This region is located between the Tropics of Cancer and Capricorn, and the world's primary producers are Brazil (2720 million kg/year), Vietnam (1650 million kg/year), and Colombia (810 million kg/year). Furthermore, the social impact of the coffee growing industry is very significant because the people who depend on this activity for all or most of their living exceeds 100,000,000 worldwide [7].

The market is divided into two groups, known as the standard and specialty markets, according to the quality of the final product, which depends on the cultivated coffee variety, the environmental conditions, and the post-harvest process. This quality is measured with a score between zero and 100 and is known as the cup quality. When the cup quality is less than 80 points, the coffee belongs to the standard market, and its selling price depends primarily on the New York Commodity Exchange. On the other hand, when the coffee has a cup quality greater than or equal to 80 points, it belongs to the specialty market, and its selling price is at least twice the standard coffee price [8]. Nevertheless, it is a fact that coffee, which is cultivated with a view toward the specialty market, needs a more careful and judicious agronomic management.

Regardless of the product's target market, coffee growers around the world face three significant challenges currently to preserve quality: (i) unpredictable climate variations, (ii) the presence of nutritional deficiencies, and (iii) attacks of pests and diseases. Concerning the latter, for instance, Coffee Leaf Rust (CLR), which is a disease considered to be the main phytosanitary problem for coffee crops, causes in Latin America losses of 30% of the efficiency of each harvest [9].

The fungus *Hemileia vastatrix* is the cause of the CLR disease, which is the major phytosanitary problem for coffee crops. Once high levels of severity are reached, the corrective actions can be minimal. Inappropriate management of the disease can harshly compromise the coffee plants, as seen in Figure 1a, resulting in only a few leaves remaining on the trees, which has a direct negative impact on the quantity and quality of the harvest [10].



**Figure 1.** Coffee Leaf Rust (CLR) effects: (a) on the Caturra variety crops; (b) on a leaf at the disease's highest development stage [11].

CLR progresses gradually in time and reaches three noticeably phases. The first one, called the “slow phase” (severity  $\leq 5\%$ ), is where the first structures responsible for the production of spores emerge and low levels of infection are evident. The second one, which is named the “fast or explosive phase” ( $5\% < \text{severity} \leq 30\%$ ), starts with the fungus sporulation and is represented by more plants getting sick in a short period. The final phase is called the “maximum or terminal phase” (severity  $> 30\%$ ) and occurs when most of the leaves are severely attacked and a small amount of healthy leaves remains. At that moment, the epidemic stops in the host due to the lack of biological matter to continue the infection. When the CLR is not controlled and the climatic conditions are favorable, the disease can develop at a daily rate of 0.19–0.38%, reduce the impact of the chemical controls, and cause significant economic damage [10].

### 1.1. Context

In the Colombian context, coffee is the most exported agricultural product, followed by cut flowers, bananas, cocoa, and sugarcane [12]. In the country, there are more than 903,000 hectares dedicated to it, and approximately 563,000 families depend directly on this economic activity. Colombian coffee has been considered one of the best soft coffees in the world, and this product has traditionally been of great importance for Colombian exports. Currently, 14,000,000 bags of 60 kg are exported every year to the USA, Japan, and Germany, among other countries [13].

In terms of employment generation and income distribution, coffee growing is a sector with superlative relevance for local economies and the maintenance of the social fabric in many regions of the country. For this reason, it is justified to contribute by solutions that strengthen the profitability of families engaged in this activity and improve their life quality, either by increasing the selling price of the product, reducing production costs, or increasing the number of units produced per unit of cultivated area.

Among the main threats for strengthening the coffee growing families' profitability, nutritional deficiencies and phytosanitary problems stand out. Phytosanitary problems are caused by pests such as the coffee borer beetle and diseases such as CLR, whose proliferation increases due to the drastic climate changes (from long drought periods to extended rainy seasons) that occur in Colombia. In the case of CLR, when the climatic conditions are unfavorable and the agronomic management deficient, at least 20% of the total expected harvest is not able to be collected. Additionally, the quality of coffee deteriorates dramatically, reducing the marketing price and increasing the costs associated with its control [10]. In extreme cases of CLR, the disease has caused devastating losses that have represented between 70% and 80% of the total harvest.

Although it is a disease with vertiginous spread and highly negative repercussions for the coffee farmers' economy, its detection and diagnosis are carried out using visual inspection while walking through the crops. This method refers to the recognition of plant diseases using visual inspections,



development scales, and standard severity diagrams for their measurement [14]. People in charge of the crops walk through them, watching and touching the plants to identify symptoms associated with the particular disease that produces them and calculate infection levels [15].

Unfortunately, because the process consists of a visual inspection, which is not done with enough regularity, most of the time, the detection of the development stage of the disease is late, its control becomes more difficult, and considerable economic losses are inevitable.

## 1.2. State-of-the-Art

Plenty of research has been done on applying technological methods and strategies to diagnose diseases [16], to detect pests [17], and to obtain nutritional information [18], among other objectives, for different types of crops. The phytosanitary status of the plantations is closely related to different crucial factors in their ecosystem, such as weather, altitude, and type of soil, among others. Therefore, several biological and engineering studies aim to implement practical solutions based on these factors to improve farming techniques to preserve healthy crops.

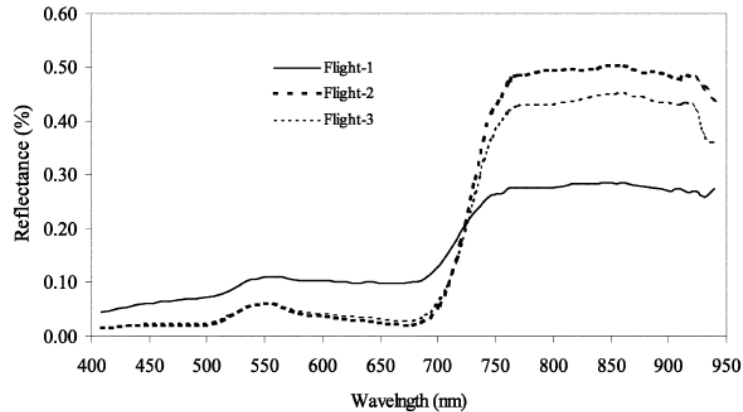
The most commonly used methods for monitoring the phytosanitary status efficiently, including those that make use of technology, are: (i) Remote Sensing (RS), (ii) visual detection, (iii) biological intervention, (iv) Wireless Sensor Networks (WSN), and (v) Machine Learning (ML) supported on a source of data. Thus, this work is intended to present recent relevant studies based on the mentioned methods for detecting anomalies on the plantations.

### (i) Remote Sensing (RS)

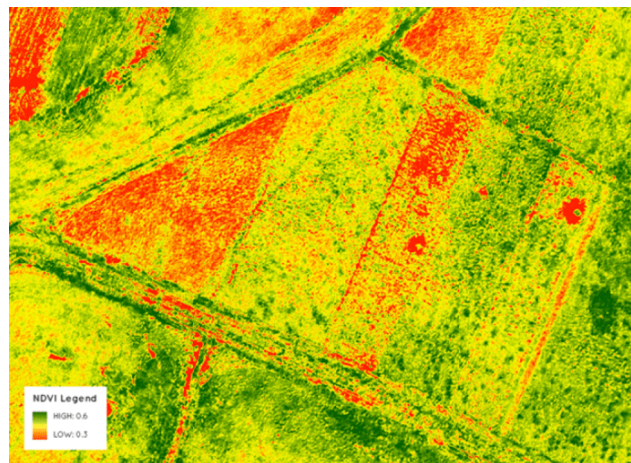
RS is based on the interaction of electromagnetic radiation with any material. In the case of agriculture, it involves the non-contact measurement of the reflected radiation from soil or plants to assess different attributes such as the Leaf Area Index (LAI), chlorophyll content, water stress, weed density, and crop nutrients, among others. Those measurements can be made using satellites, aircraft, drones, tractors, and hand held sensors [19]. In addition to measuring reflected radiation, there are two other RS techniques that analyze fluorescent and thermal energy emitted by the leaves. However, the most common technique is reflectance, because the amount of reflected radiation from the plants is inversely related to the radiation absorbed by their pigments, and this can serve as an indicator of their health status [19]. RS helps the indirect detection of problems in agricultural fields since this method captures unusual behaviors in crops' reflectance, which can be caused by factors like nutritional deficiencies, pests and diseases, and water stress. In 2017, Calvario et al. [20] monitored agave crops using Unmanned Aerial Vehicles (UAVs) and integrating RS with unsupervised machine learning (*k*-means) to classify agave plants and weed. In 2003, Goel et al. [21] studied the detection of changes in the spectral response in corn (*Zea mays*) due to nitrogen application rates and weed control. For that purpose, the researchers employed a hyperspectral sensor called the Compact Airborne Spectrographic Imager (CASI) and analyzed the reflectance values of 72 bands with a wavelength between 409 and 947 nm, which comprise part of the visible and Near-Infrared (NIR) regions of the electromagnetic spectrum. The obtained results demonstrated the potential of detecting weed infestations and nitrogen stress using the hyperspectral sensor CASI. Specifically, the researchers found that the best fitting bands for the detection were the wavelength regions near 498 nm and 671 nm, respectively, as seen in Figure 2.

It has been shown that using satellites' multispectral images, it is possible to detect the location of crops [22], but the resolution of satellites images does not allow early detection of the phytosanitary of individual lots of plants. Regarding the phytosanitary status of the plants, the water and the type of soil are two components that play an essential role in their health. In 2017, Bolaños et al. [23] proposed a characterization method using the visible and infrared spectrum to identify these components, through low cost cameras with two different filters, Roscolux #19 and Roscolux #2007, and a multi-rotor air vehicle. Through this method and using portable and highly qualified devices, those hard-to-reach places were monitored and analyzed to detect anomalies that may cause diseases in the

crop. This monitored phase provided a characterization of the Normalized Difference Vegetation Index (NDVI), as seen in the example of Figure 3, which was used to categorize essential characteristics of the crop, such as crop health, diseased plants or soil, and water or others.



**Figure 2.** Reflectance (%) of the corn response during different flights under normal nitrogen rates and no weed control [21], Copyright Elsevier, 2003.



**Figure 3.** Characterization of the NDVI with low cost solutions [24].

In 2017, Chemura et al. proposed a method to predict the presence of diseases and pests early among coffee trees based on unnoticeable water stress. For that purpose, multispectral scanners with filters with wavebands from the visual spectrum and near infrared region were placed on a UAV [25]. The wavebands scanner results showed inflections points between the regions 430 nm and 705–735 nm due to the water content in coffee trees. These results underlined the importance of a suitable irrigation plan according to the water requirements of the trees, causing an improvement in productivity. Although the later region indicated relevant values, the waveband of 430 nm was the most relevant band of remote sensing for predicting the water plant content directly related to its stress. However, in [25], the authors remarked that although the results were promising, there were some missing valid components that could allow the model to be suitable and testable in real conditions. For that purpose, they recommended using hyperspectral cameras, which provide more precise measured waveband results.

(ii) Visual Detection

The detection of visual symptoms uses the changes in the plant’s appearance (colors, forms, lesions, spots) as an indicator of it being attacked by a disease or pest [15]. In the survey of Hamuda et al. [26], image based plant segmentation, which is the process of classifying an image into plant and non-plant, was used for detecting diseases in plants [27]. For instance, for the evaluation of the CLR’s infection percentage in a specific lot, the number of diseased leaves in 60 random trees had to be divided into the total number of leaves in those trees and multiplied by 100 (see Equation (1)). A leaf is considered diseased with CLR when chlorotic spots or orange dust are observed on it. The severity of the disease can be divided into five categories depending on the number and diameter of rust orange spots, as seen in Figure 4.

$$Average\ infection\ \% \ in\ the\ lot = \frac{Number\ of\ diseased\ leaves\ in\ the\ 60\ trees}{Total\ number\ of\ leaves\ in\ the\ 60\ trees} \times 100 \quad (1)$$

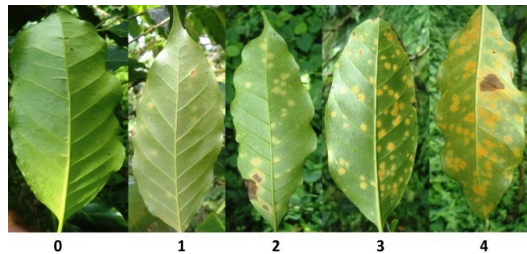


Figure 4. CLR development stages [28].

A visual inspection can be carried out to detect the presence of chlorotic spots on the leaves, which are then used for measuring the incidence and severity of the disease [10].

To understand the conditions conducive to the development of CLR and, subsequently, refine the disease control, Avelino et al. [29] monitored such development on 73 coffee crops in Honduras for 1–3 years. Thereby, through the analysis of production situation variables such as climate, soil components, coffee tree productive characteristics, and crop management patterns, the researchers aimed to establish a relationship with the presence of rust. The result of this research indicated that CLR epidemics depend on the diverse production situations based on Table 1, linked as well to the local conditions of the plantation. Due to the above, these results reflect the need for the consideration of a certified growing system that aims for sustainability, taking into consideration production situations and, thus, preventing the development of pests and diseases.

Table 1. Kinds of variables that describe the importance of coffee plots in the presence of CLR [29].

Kind of Variable	Relevance
Climate variation (Altitude and rainfall)	High
Soil components	Medium–low
Cropping practices	Medium
Coffee tree productive characteristics	High

(iii) Biological Intervention

Several authors stated the importance of the relationship between living beings sharing the same environment. One of them was Haddad et al. [30], who in 2009, proposed a study to determine if seven selected isolated bacteria under greenhouse conditions would efficiently detect and control CLR. For the development of this research, they inoculated these bacteria: six *Bacillus* sp., B10, B25, B157, B175, B205, and B281, and one *Pseudomonas* sp., P286, which help to detect and control CLR in

the early development stages, according to a preliminary result presented by Haddad et al. (2007). For the experiment, two important coffee varieties, Mundo Novo and Catuai, were selected due to the high susceptibility to CLR. Therefore, for three years, the varieties with the disease interacted with different treatments (bacteria) to analyze the behavior evolution between them. Based on the results of the treatments, the isolates P286 and B157 were as efficient as the copper fungicide in controlling the rust. Hence, considering the harmful effects due to the copper fungicide, the application of biological control with the B157 isolate of *Bacillus* sp. may be a reliable alternative solution to CLR management. That is why this research displayed the opportunity to successfully biocontrol CLR, for specialty coffee growers.

Jackson et al. [31], in 2012, proposed as well a biological detection and control based on a fungus, *Lecanicillium lecanii*. Their primary interest in the crops, in general, was the analogy of the coexistence of organisms in a specific environment with defined conditions that encounter a perfect balance. Given the above, the biological control system of the *A. instabilis* ants were mutualistically associated with the white halos of the fungus, *Lecanicillium lecanii*, based on the CLR effect.

However, the hypothesis stated the possibility that spores from *Lecanicillium lecanii* help to attack the *Hemileia vastatrix* before the rainy season. The effect of the time delay of *Lecanicillium lecanii* in *Hemileia vastatrix* resulted in a relationship between the two fungi and the ants not to be demonstrated, in spite of the control experiment resembling the real world. In conclusion, the restriction of biotic factors directly affects the development of CLR; therefore, for future work, it is important to consider the climate variation of an ecosystem to be able to predict such development [31].

#### (iv) Wireless Sensor Networks (WSN)

Wireless Sensor Networks (WSN) are a technology that is being used in many countries worldwide to monitor different agricultural characteristics in real time and remotely. It consists of multiple non-assisted embedded devices, called sensor nodes, that collect data in the field and communicate them wirelessly to a centralized processing station, which is known as the Base Station (BS). The BS has data storage, data processing, and data fusion capabilities, and it is in charge of transmitting the received data to the Internet to present them to an end-user [32]. Once the collected data are stored on a central server on the Internet, further analysis, processing, and visualization techniques are applied to extract valuable information and hidden correlations, which can help to detect changes in crop characteristics. These changes could be used as indicators of phytosanitary problems such as nutritional deficiencies, pests, diseases, and water stress. WSN is a powerful technology since the information of remote and inaccessible physical environments can be easily accessed through the Internet, with the help of the cooperative and constant monitoring of multiple sensors [33]. The sensor nodes in a WSN setup can vary in terms of their functions. Some of them can serve as simple data collectors that monitor a single physical phenomenon, while more powerful nodes may also perform more complex processing and aggregation operations. Some sensors can even have GPS modules that help them determine their particular location with high accuracy [33]. The most common sensors used in WSN for agriculture are the ones that collect climate data, images, and frequencies. Chaudhary et al. [34] emphasized in 2011 the importance of WSN in the field of PA by monitoring and controlling different critical parameters in a greenhouse through a microcontroller technology called Programmable System on a Chip (PSoC). As a consequence of the disproportionate rainfall dynamics, the need for controlling a suitable water distribution meeting those parameters inside the greenhouse arises. Thereby, the study tested the integration of wireless sensor node structures, with high bandwidth spectrum telecommunication technology. Mainly, it was proven that the integration was useful to determine an ideal irrigation plan that met the specific needs of a crop based on the interaction of the nodes within the greenhouse. Furthermore, the researchers recommended using reliable hardware with low current consumption to develop WSN projects, because it generates more confidence for the farmers concerning its incorporation with their crops and provides a longer battery life.

Besides, Piamonte et al. [35] proposed in 2017 a WSN prototype for monitoring the bud rot of the African oil palm. With the use of pH, humidity, temperature, and luminosity sensors, they aimed to measure climate variations and edaphic (related to the soil) factors to detect the presence of the fungus that causes the disease indirectly.

#### (v) Machine Learning

The domain concerned with building intelligent machines that can perform specific tasks just like a human is called Artificial Intelligence (AI) [36]. One of the main subareas of AI is Machine Learning (ML), which aims to extract complex patterns from large amounts of raw data automatically to predict future behaviors. When the extracting process of those patterns is taken to a more detailed level, where computers learn complicated real-world concepts by building them out of simpler ones in a hierarchical way, ML enters one of its most relevant subsets: Deep Learning (DL) [37]. The functionality of DL is an attempt to mimic the activity in layers of neurons in the human brain. The central structure that DL uses is called an Artificial Neural Network (ANN), which is composed of multiple layers of neurons and weighted connections between them. The neurons are excitable units that transform information, whereas the connections are in charge of rescaling the output of one layer of neurons and transmitting it to the next one to serve as its input [38]. Inputting data such as images, videos, sound, and text through the ANN, DL builds hierarchical structures and levels of representation and abstraction that enable the identification of underlying patterns [36]. One application of finding patterns through DL can be for estimating plant characteristics using non-invasive methodologies by means of digital images and machine learning. Sulisty et al. [39] presented a computational intelligence vision sensing approach that estimated nutrient content in wheat leaves. This approach analyzed color features of the leaves' images captured in the field with different lighting conditions to estimate nitrogen content in wheat leaves. Another work of Sulisty et al. [40] proposed a method to detect nitrogen content in wheat leaves by using color constancy with neural networks' fusion and a genetic algorithm that normalized plant images due to different sunlight intensities. Sulisty et al. [41] also developed a method for extracting statistical features from wheat plant images, more specifically to estimate the nitrogen content in real context environments that can have variations in light intensities. This work provided a robust method for image segmentation using deep layer multilayer perceptron to remove complex backgrounds and used genetic algorithms to fine tune the color normalization. The output of the system after image segmentation and color normalization was then used as an input to several standard multi-layer perceptrons with different hidden layer nodes, which then combined their outputs using a simple and weighted averaging method. Fuentes et al. [42] presented a robust deep learning based detector to classify in real-time different types of diseases and pests in tomatoes. For such a task, the detector used images from RGB cameras (multiple resolutions and different devices such as mobile phones or digital cameras). This method detected if the crop had a disease or pest and which type it was. Similarly, Picon et al. [43] developed an automatic deep residual neural network algorithm to detect multiple plant diseases in real time, using mobile devices' cameras as the input source. The algorithm was capable of detecting three types of diseases on wheat crops: (i) *Septoria* (*Septoria tritici*), (ii) tan spot (*Drechslera tritici-repentis*), and (iii) rust (*Puccinia striiformis* and *Puccinia recondita*). Related to CLR, research has been done, such as that by Chemura et al. [44], who evaluated the potential of Sentinel-2 bands to detect the CLR infection levels early due to its devastating rates. Through the employment of the Random Forest (RF) and Partial Least Squares Discriminant Analysis (PLS-DA) algorithms, such levels could be identified for early CLR management. The researchers employed the variety of Yellow Catuai, which was chosen due to its CLR susceptibility. In this matter, Chemura et al. considered only seven Sentinel-2 Multispectral Instrument (MSI) bands due to the high resolution stated by previous works in biological studies. Based on the selected bands, the research results determined that the CLR reflectance was higher in NIR regions of the spectrum, as could be seen in leaves from the bands B4 (665 nm), B5 (705 nm), and B6 (740 nm). These bands achieved a high overall CLR discrimination of 28.5% and 71.4% using the RF and PLS-DA algorithms respectively.

Thus, the band and vegetation indices derived from the MSI of Sentinel-2 achieved the detection of the disease and an evaluation of CLR in the early stages, avoiding unnecessary chemical protection in healthy trees.

In 2017, Chemura et al. [45] studied the detection of CLR through the reflectance of the leaves at specific electromagnetic wavelengths. The objective of their investigation was to assess the utility of the wavebands used by the Sentinel-2 Multispectral Imager in detection models. The models were created using Partial Least Squares Regression (PLSR) and the non-linear Radial Basis Function partial Least Squares Regression (RBF-PLS) machine learning algorithm. Then, both models were compared, resulting in a low accuracy prediction of the state of the disease for the PLSR, due to its over-fitting, and a high accuracy prediction for the RBS-PLS model. Additionally, Chemura et al., through weighting of the importance of the variables, found that the blue, red, and RE1 bands had a high model correlation, but the implementation excluding the remaining four bands led to lower accuracy in both models. On the other hand, if more than one NIR or red edge (RE) band were used, then the RBS-PLS model developed would over-fit, resulting in a non-transferable model. However, Chemura et al. emphasized the utilization of the RBS-PLS model due to its machine learning advantage and its excellent adaptation to possible model over-fitting.

### 1.3. Conclusions of the Literature Review

The presented state-of-the-art showed that several researchers sought the detection of any vital element like water stress, nitrogen levels, and vegetation indexes that could lead to an improvement of production and quality in crops, which translated to an increase in profitability. However, most of the research did not integrate different means of detecting CLR to have more insights and better accuracy in predicting this disease. Furthermore, the determination of the infection percentage of the crop through visual inspection is a tedious task, which is also laborious, time consuming, and subject to human error and inconsistency [46]. For this reason, the objective of this research is to evaluate to what extent it is possible to diagnose the CLR development stage in the Colombian Caturra variety (the most susceptible to the disease) through a technological integration system that involves Remote Sensing (RS), Wireless Sensor Networks (WSN), and Deep Learning (DL). Adequate management of CLR could preserve the quality and selling price of the final product, reduce production costs by rationing control costs, and protect productivity. The present research aims to facilitate the management of the most dangerous disease in the Colombian Caturra variety's coffee production to strengthen the profitability of the rural inhabitants.

The present work provides empirical evidence of a novel diagnostic method for the classification of the development stage of CLR in coffee crops, by means of a technological integration of image data (RS), WSN, and DL. This contribution allows coffee growers to detect CLR disease automatically, thus optimizing the production and maintenance of their crops and replacing the task of manual inspection. Through this method, the performance evaluation is done, and the results are presented to conclude to what extent it is possible to diagnose CLR disease. Thus, this information can be useful for coffee growers to determine if the integration of RS, WSN, and DL in our method could positively impact their profitability.

## 2. Proposed Method

The design of experiment implemented in this research was a Completely Randomized Design (CRD). It was used to compare two or more treatments considering only two sources of variability: treatments and random error. The objective of using this design of experiment in this project was to analyze whether the diagnosis of the CLR development stage through the integration of RS, WSN, and DL was similar to the one made with a traditional visual inspection. A summarized diagram of this process is shown in Figure 5.

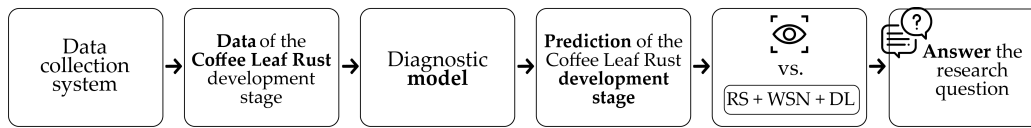


Figure 5. Proposed methodology flowchart (based on [44]).

In that sense, the study factor was the type of inspection, which had two levels (“visual inspection” and “technological integration”), and the response variable was the development stage of the disease, which was a whole number between 0 and 4. Thereby, the fundamental hypothesis to prove, presented in Equation (2), helped by deciding whether Treatments 1 (“visual inspection”) and 2 (“technological integration”) were statistically equivalent with respect to their means [47].

$$\begin{aligned}
 H_0 : \mu_1 &= \mu_2 \\
 H_A : \mu_1 &\neq \mu_2
 \end{aligned}
 \tag{2}$$

The procedure for proving the mentioned hypothesis is called Analysis Of Variance (ANOVA) and required a data table containing a row for each observation and a column for each treatment indicating the measurements of the response variable. This procedure separated the variability due to the treatments from the one attributed to the random error and compared them. If the former was higher than the latter, the means of the treatments were different, and thus, the type of diagnosis influenced the determined CLR development stage. Otherwise, the means were statistically equivalent, and it was possible to conclude that the visual inspection and the technological integration were similar for diagnosing the disease. Lastly, it is essential to mention that the significance level that was used for proving the hypothesis was 10% ( $\alpha = 0.1$ ), since the problem at hand was related to agriculture, where many noise factors associated with the variation of environmental conditions were involved [47].

For the data collection experiment, 16 six month old, healthy coffee plants coming from Jardín, Antioquia, were used. Those plants were stored in a Universidad EAFIT’s greenhouse. A biology team was in charge of their transplantation, agronomic management (elimination of weeds, fertilization, and fumigation), inoculation, and supervision. For the inoculation, the biology team followed the process described in Chemura et al. [44]. It is relevant to clarify that a new group of diseased plants was held as a reserve in case the inoculation of the healthy plants did not take effect over time.

Furthermore, an engineering team was dedicated to the design and assembly of a system, in the same greenhouse, that integrated RS and WSN. It allowed building a scale crop, recording different characteristics of it regularly, and storing them on a remote server to analyze its phytosanitary status later using DL. In that way, once the plants were inoculated and the system was verified, they were transplanted to it so that the data collection may begin. For that purpose, the scale crop was divided into four lots with certain differences in their agronomic management, which sought to recreate various circumstances of a real coffee crop. Thereby, a greater number of scenarios were covered, and the false positive rate regarding the diagnosis was reduced. *LOT 1* contained four non-inoculated plants, and they were neither fertilized nor fumigated; *LOT 2* had four non-inoculated plants and was fertilized but not fumigated; *LOT 3* had four inoculated plants, and they were also fertilized but not fumigated; and *LOT 4* had four inoculated plants, and they were neither fertilized nor fumigated. The previous distribution can be seen in Figure 6.

Finally, the visual inspections for diagnosis of the CLR development stage were carried out by the biology team for three months. Once per day, one of them examined the severity of the disease for each lot and indicated the value of the response variable for each observation; this measure corresponded to the ground truth. Similarly, the technological system automatically recorded the scale crop’s characteristics from each lot seven times per day at different moments (with and without sunlight, because the field sensors and cameras had different illuminance requirements), assigning to each of these samples the above mentioned daily ground truth. After the data collection phase

finished, the diagnostic model using DL was generated, and a comparative data table for the statistical analysis was produced, based on its predictions and the results of the visual inspections. As it was expected that a considerable amount of observations would be made, only 25% of all collected data were used for the statistical study. It should also be noted that, as was recommended, the order of the table’s entries were randomized before executing the analysis in order to minimize bias.

<i>LOT 2</i>	<i>LOT 1</i>
No inoculation	No inoculation
Fertilization	No fertilization
No fumigation	No fumigation
Inoculation	Inoculation
Fertilization	No fertilization
No fumigation	No fumigation
<i>LOT 3</i>	<i>LOT 4</i>

Figure 6. Data collection distribution.

2.1. Experimental Testbed

To evaluate to what extent it was possible to diagnose the CLR development stage in the Colombian Caturra variety through the integration of RS, WSN, and DL, it was necessary to obtain empirical evidence employing an experiment. Therefore, an experimental testbed prototype was built, which included a scale coffee crop. This testbed was capable of simulating different agronomic conditions and allowed capturing data for diagnosing the disease. The experimental testbed consisted of a data collection system prototype that integrated remote sensing and wireless sensor networks. In this testbed, the coffee plants were grouped, combined with the soil, and then divided into four lots. Furthermore, they were separated to inoculate CLR in half of them, and after that, the four lots were assembled again. For their agronomic management, fertilizer and fungicide were distributed and incorporated. Then, each lot was isolated from the others to make the four lots independent, and the whole scale crop was combined with a rain emulation system and a wind system. Both rainfall and wind speed for the whole crop were perceived. Furthermore, using sensors in each lot, pH, illuminance, temperature, humidity, and electrical conductivity were perceived, which will be further called “sensor data”, and RGB and multispectral images were captured. RGB pictures were acquired through a regular RGB camera with a resolution of 720 p. These cameras were positioned on the bottom of the plants since CLR was commonly visible at the underside of the leaf [10]. Regarding the multispectral cameras, which allowed capturing the reflected radiation of wavelengths that were not perceptible to the human eye, two cameras from MAPIR®, called Survey3, were used. Based on the information cited in the state-of-the-art [21,23,25], the Red + Green + NIR (RGN) and Red Edge (RE) camera filters were chosen as being suitable to identify crop diseases, including CLR. Thus, one camera centered in the wavelengths 660 nm–550 nm–850 nm and another one centered in the 735 nm wavelength were selected to capture images from the top of the plants. The Survey3 incorporated a Sony® Exmor R IMX117 12MP sensor and a sharp non-fish eye lens for perceiving light in specific wavelengths. The created experimental testbed is shown in Figure 7.



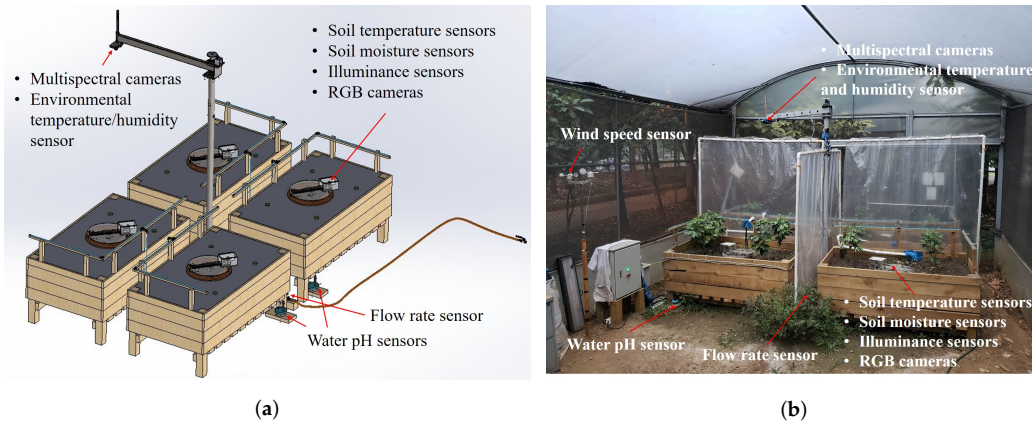


Figure 7. Experimental testbed: (a) 3D CAD model; (b) implemented prototype.

Afterwards, the state of each lot was integrated with the expert’s visual inspection information to diagnose the CLR development stage, and then, this information was clustered with the collected data. To finish the data collection process, data were stored locally and sent to a remote server over the Internet.

On the other hand, the data that were received on the remote server were preprocessed for cleaning purposes and stored in a remote database. An example of the LOT 3 directory’s content on the remote server after one data collection routine was concluded is presented in Figure 8.



Figure 8. LOT 3 directory’s content after a data collection routine.

To clarify how a data collection routine worked, Figure 9 details the whole pipeline from the sensor readings and image captures until the remote storage. The data from sensors were gathered and smoothed by a microcontroller. RGB and multispectral images were captured by the cameras. The totality of the data was collected by a Single Board Computer (SBC), which continually notified the progress to the Internet of Things (IoT) platform (see Figure A3 inside Appendix C for the IoT platform dashboard) while it created a single data package. The package containing the documents with the lots and general data, as well as the images was stored locally. Furthermore, the documents were inserted into the remote MongoDB®, which resided in the data center, and the entire data package was uploaded via Secure File Transfer Protocol (SFTP) to the data center’s file system. At that point, the data collection routine finished.

Finally, it is also relevant to mention how the collected data can be reviewed so that the process can be verified. Using a personal computer, the IoT platform, the single board computer, and the data center could be accessed over the Internet. The access to the IoT platform required a web browser, while the single board computer and the data center could be remotely inspected through the graphical desktop sharing system Virtual Network Computing (VNC) or the cryptographic network protocol Secure Shell (SSH).

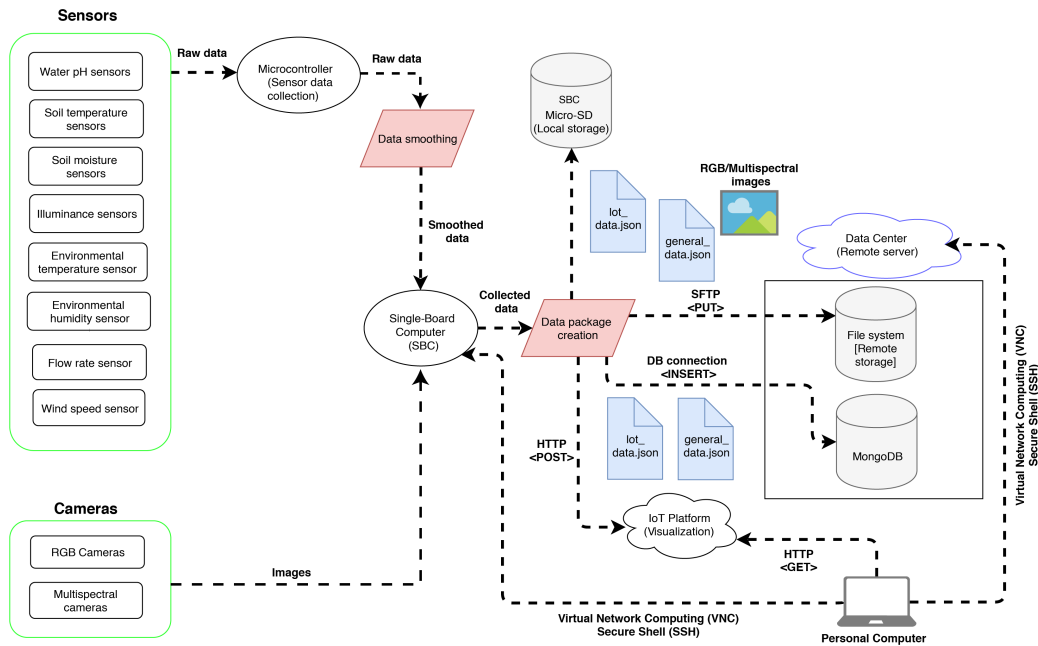


Figure 9. Data collection pipeline.

## 2.2. Machine Learning Pipeline

To create an adequate model for diagnosing the CLR development stage, the stored data were first divided into two sets, namely training (with cross-validation) and test. The training set was processed to build the diagnostic model with cross-validation, which served to assess its intermediate performance and tune it. Once the diagnostic model was generated, the test set was used for evaluating its final performance. All the developed models and cloud storage were implemented using an academic data center.

Within the framework of this project, the data center was used to store the data collected remotely on the physical part of the prototype. Both the MongoDB<sup>®</sup> instance in it, as well as its file system made the replication of single-board computer’s local storage possible and facilitated the ubiquitous access to that information. Furthermore, the data center was the place where the data preprocessing, model generation and CLR development stage diagnosis occurred. It is also relevant to mention that the software libraries used for the implementation were Python 3.6.0, NumPy 1.16.0 [48], Pandas 0.24.0 [49], Scikit-learn 0.20.2 [50], and Keras 2.2.4 [51] running on top of TensorFlow 1.12 [52].

The machine learning pipeline model to show how the collected data were manipulated to extract the model that was used to diagnose the development stage of the disease in question is shown in Figure 10.

This pipeline model initially consisted of four sub-directories ranging from *LOT 1* to *LOT 4* where each lot’s data would be correspondingly labeled later on. For that purpose, the biology team determined the labels by carrying out visual inspections in the field on all plants once a day during the whole data collection phase. In that sense, it assigned a whole number between 0 and 4 to each plant on each lot, evaluating the plant leaves’ severity level, and calculated the specific lot’s label as the rounded average of its four plants’ disease development stages. All data directories of the current day and corresponding lot were labeled with the value of the last visual inspection, which was determined in the most recent checkup.

Subsequently, a new *rgb\_images* directory containing five sub-directories (ranging from 0 to 4) representing the diseases’ five stages was created. In these five sub-directories, RGB images coming from all lots (*LOT 1* to *LOT 4*) were correspondingly stored according to their label. Similarly, the sensor

data, which were stored as a JavaScript Object Notation File (JSON), and the multispectral images had the same label as the RGB images belonging to the same lot. Furthermore, in the case of images in general, they were visually checked one-by-one to keep only the ones with valuable content (focus, brightness level) and remove the others. In addition to this, a script was executed to eliminate the irrelevant JSON files (those with missing values and outliers), as well as the sub-directories that ended up with no content. The last two actions were part of the depuration stage. In the end, five sub-directories would exist containing the data from all lots (LOT 1 to LOT 4) adequately labeled. Those sub-directories were the ones that were used for the generation and final evaluation of the diagnostic model, taking into account that the diagnosis occurred at the lot level.

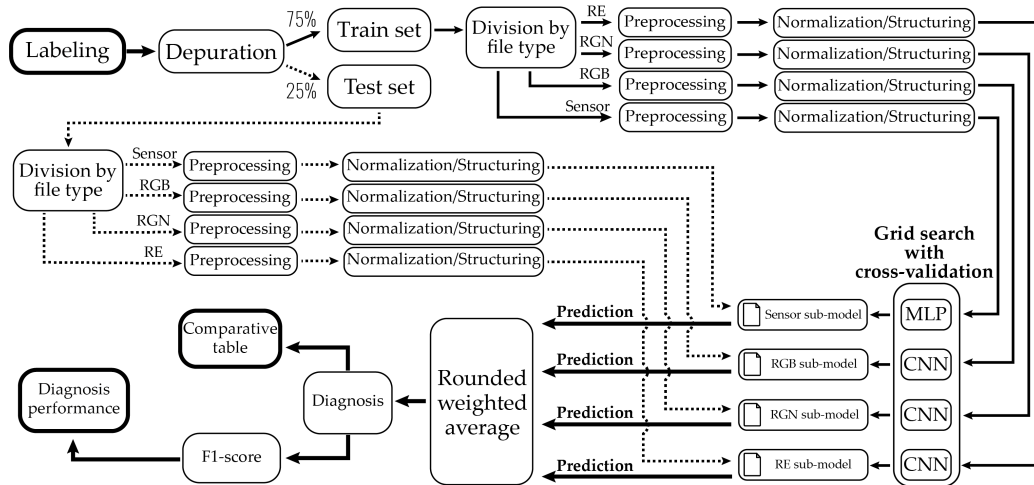


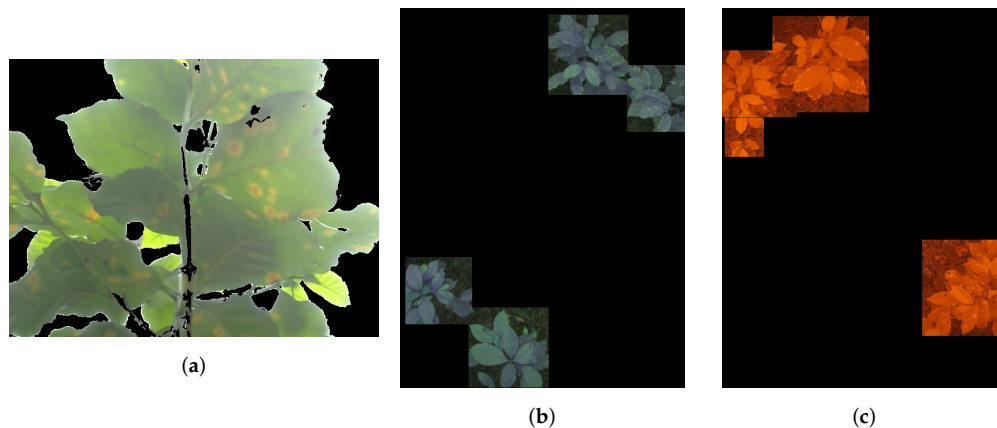
Figure 10. Machine learning summarized pipeline model.

Once all the data were correctly distributed, the content of each of the five sub-directories was virtually shuffled, and the elements per file type were counted for every sub-directory. Then, per label sub-directory, the minimum of those values was found. Twenty-five percent was calculated, and the file type associated with that minimum was determined. The resulting numbers indicated the number of files per respective determined file type and per label that could be used, at most, for testing the diagnostic model. Taking those threshold numbers into account, the shuffled lot data directories within each label subdirectory were individually analyzed to split them into two groups, namely training and test sets. If a particular lot/s data directory was considered as complete (i.e., it had a JSON file, the two multispectral images, and at least one RGB image) and supposing that using its files for testing did not exceed the corresponding threshold, then it was copied under the same structure to another location in order to feed the test set. Otherwise, the lot data directory was also copied, but to grow the training set. Thereby, the training set (~75% of all data) was used to train and tweak the model, while the test set (~25% of all data), with no overlapping with training set, was only incorporated at the time of the diagnosis evaluation. The data distribution after the above mentioned process was importantly imbalanced, as seen in Table 2. It can be noted that Stage 1 was not included in the table. This was due to the fact that only one sample was identified in that stage. Consequently, it could not be used for the model construction, and therefore, it could be considered as not relevant.

Table 2. Data distribution between the training and test sets by each CLR development stage.

# of Samples	Stage 0	Stage 2	Stage 3	Stage 4	Total
Training	711	55	90	112	968
Test	149	12	18	23	202

After the two sets were correctly obtained, one submodel was generated for each file type, i.e., sensor data (JSON), RGB, RGN, and RE. For the JSON files, Multi-Layer Perceptron (MLP) was used, whereas Convolutional Neural Networks (CNNs) were implemented to classify the RGB, RGN, and RE images. For that purpose, the data in the training set were first divided into four subdirectories according to the file type, while preserving the same structure. Then, each of them was preprocessed so that the noise was removed from the images, and the irrelevant keys in the documents were also identified and eliminated. Figure 11 illustrates an example of preprocessed image files.



**Figure 11.** Example of preprocessed image files: (a) RGB image; (b) RGN image; (c) RE image.

After that, the corresponding data were loaded within each submodel's generation, divided into feature data (the files themselves) and label data (the names of the label subdirectories that contained the files), and permuted. Thereby, the data were randomly mixed while it was still possible to know each feature's respective label unequivocally. Then, if applicable, the data were normalized and structured to scale the input and format, as was recommended when using deep ANNs. The normalization used for this experiment was the z-score (subtracting the mean of the feature and dividing by its standard deviation), which scaled the data to have the properties of a standard normal distribution [53]. Upon having the data prepared, different architectures and hyperparameter values were tried to train the submodel to tune it to reach higher performance values on the predictions.

The technique used for tuning the submodel is called grid search with cross-validation. It consisted of executing an exhaustive search over specified hyperparameter values for an estimator to find out which combination achieved the best performance, which was by default the higher accuracy, but different metrics could be chosen. One candidate estimator for each combination of hyperparameters was built and evaluated, so that the best estimator, its attributes, and its average performance could be extracted once the search was complete [54]. Furthermore, the procedure for measuring the average performance of each candidate estimator during the generation of the submodel is called *k*-Fold Cross-Validation, where *k* separate learning experiments are run on the the same estimator to calculate *k* performance values and average them. To achieve this, the feature and label data were split at the beginning into *k* non-overlapping subsets (also known as "folds"), so that for every experiment, one different fold was kept for measuring the performance, whereas the remaining *k* - 1 were put together to form the training set to fit the estimator [55]. Finally, when the grid search processes concluded, the four submodels were extracted and saved for the definitive diagnosis about the CLR development stage.

To select the best estimator during the grid search with cross-validation, the chosen metric was the  $F_1$ -score, which, in the multi-label case, was the weighted average of the labels'  $F_1$ -scores. This metric was used due to the importantly imbalanced dataset (skewed classes) between the development stages of the CLR, as seen in Table 2. The  $F_1$ -score of the label *L* is a value in the [0, 1] range, and it was calculated as the harmonic mean of the estimator's precision and recall with respect to *L* (see

Equation (3)). The precision with respect to  $L$  is the ratio of the number of times that  $L$  was correctly predicted to the overall number of times that  $L$  was predicted. Furthermore, the recall with respect to  $L$  is the ratio of the number of times that  $L$  was correctly predicted to the overall number of times that  $L$  should have been predicted. Thereby, the general  $F_1$ -score reaches its best value at 1, indicating that the estimator perfectly matched reality, and its worst at 0, showing that the estimator never coincided with reality [53].

$$F_1\text{-score}_L = \frac{2 * precision_L * recall_L}{precision_L + recall_L} \quad (3)$$

At this point, the data that were kept to be only incorporated at the time of the diagnosis evaluation were brought up. First, the submodels were loaded. Then, each lot's data directory contained in the test set was submitted to the following process. At the beginning, its data were divided according to the file type. After that, each type was sent to its corresponding submodel, where it was first cleaned, normalized, and structured, applying the same particular procedures that were used to prepare the data for the submodel generation. Subsequently, the submodel made its prediction based on the trained model output. It is also relevant to mention that, considering that the diagnosis was made at the lot level, the RGB submodel could be used up to four times per lot data directory before retrieving its result (which was the rounded average of its predictions). The final step consisted of combining the outcomes of the four submodels and calculating their rounded weighted average, the weights being the respective  $F_1$ -scores. Thereby, the definitive lot's CLR diagnosis was obtained, and it was recorded along with the processed lot's data directory label. Once the whole test set was covered, a table showing comparative results was generated for the statistical analysis, and the performance reached by the composite model was assessed with the calculation of the  $F_1$ -score. Figure A2 from Appendix B illustrates the above machine learning pipeline in a detailed manner, and Table 3 shows the selected hyperparameters and obtained  $F_1$ -score for each of them. Tables A1, A2 and A3 from Appendix D details the architectures of the submodels.

**Table 3.** Hyperparameters and  $F_1$ -score for each generated submodel.

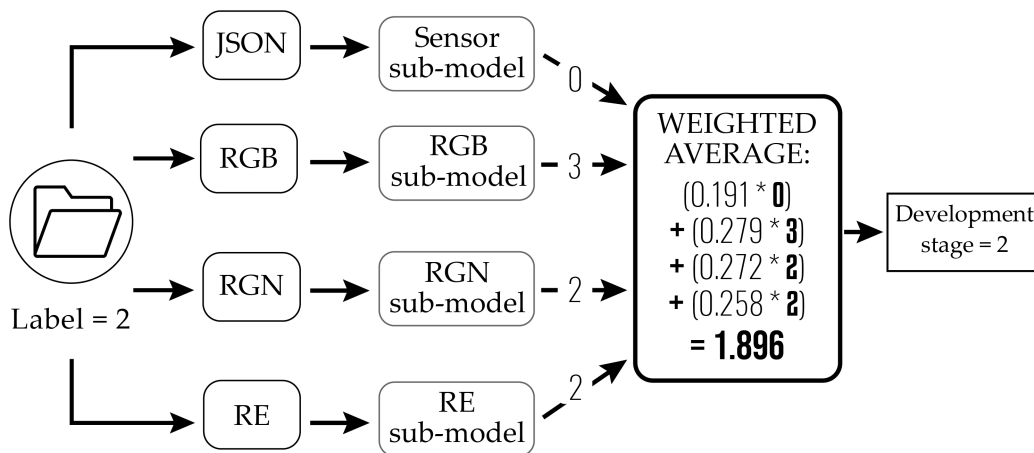
Submodel	Batch Size	Epochs	Kernel Initializer	Activation	Rate	Optimizer	$F_1$ -Score (Cross-Val Set)
Sensor data	16	20	normal	ReLU	0.4	Adam	0.651
RGB	16	6	glorot_uniform	ReLU	0.4	Adam	0.949
RGN	32	9	glorot_uniform	elu	0.3	Adam	0.928
RE	16	6	normal	ReLU	0.4	Adam	0.878

The last step of the proposed ML pipeline consisted of integrating the four presented submodels and evaluating the composite model, i.e., diagnosing the CLR development stage through it, creating a comparative table with the results achieved and calculating the model's performance. For that purpose, a model evaluator script was implemented. This script was in charge of loading the submodels into memory, iterating over the whole test set, taking each lot data directory within it, dividing the contained files according to their type and preprocessing them, resizing them to reduce the spatial complexity (in the case of images), normalizing and structuring each file according to the submodels' expected input, and sending them to their corresponding submodel to get a prediction. In addition, the script allowed gathering the four predicted labels and calculating their rounded weighted average, since the generated submodels presented different performances for diagnosing the CLR development stage. Table 4 shows the weights for the predictions of each submodel, which were determined as the ratio of each  $F_1$ -score in Table 3 with respect to the sum of all  $F_1$ -scores.

**Table 4.** Weights for the predictions of each submodel.

Submodel	Weight for Predictions
Sensor Data (JSON)	0.191
RGB	0.279
RGN	0.272
RE	0.258

To further explain the weighted average, let us assume that a sample folder with all the collected data (sensor data, RGB, RGN, and RE images) was labeled as CLR Development Stage 2 (*Label = 2*). Then, these data inside this folder were fed into the developed submodels (sensor data, RGB, RGN, and RE submodels) which produced an output based on their trained model. Let us assume that the sensor data submodel classified this as 0, the RGB submodel as 3, the RGN submodel as 2, and the RE submodel as 2. Then, considering the weights from Table 4, the averaged development stage would be approximately 1.90. Then, rounding this value up, the final output of the ML pipeline would be *DevelopmentStage = 2*. This example is shown in Figure 12.



**Figure 12.** Machine learning classification example.

### 3. Results

The results of this experiment were a composite trained model with an  $F_1$ -score of 0.775. This model was tested using ANOVA to prove the validity of the hypothesis presented in Section 2, with respect to the visual inspection and our proposal using the technological integration methods. The  $p$ -value obtained was 0.231, which was greater than the significance  $\alpha = 0.1$ . This result indicated that the proposed method for automatically detecting the CLR disease presented an equivalent performance compared to the manual/visual inspection method (the ANOVA test will be further discussed in Section 3.1). All the inputs for the obtained results are detailed below.

On the one hand, it must be mentioned that, during the data collection phase, the biology team had to replace 12 coffee plants of the scale crop with external diseased ones because the inoculation did not take effect after two months (all plants stayed in Development Stage 0).

On the other hand, the training set used for fitting the submodels was composed of 968 directories. In total, they contained 672 sensor data (JSON) files, 2192 RGB files, 603 RGN files, and 641 RE files. In addition, the test set employed for the composite model evaluation comprised 202 lot data directories, with 224 sensor data (JSON), 730 RGB files, 202 RGN files, and 202 RE files. Finally, after evaluating the diagnosis of the CLR development stage in the Colombian Caturra variety employing the created DL model, a comparative table, along with a performance table, was successfully generated. Figure A1 from Appendix A shows the comparative table for the statistical analysis. Table 5 presents the definitive  $F_1$ -score reached by each submodel and the composite model.

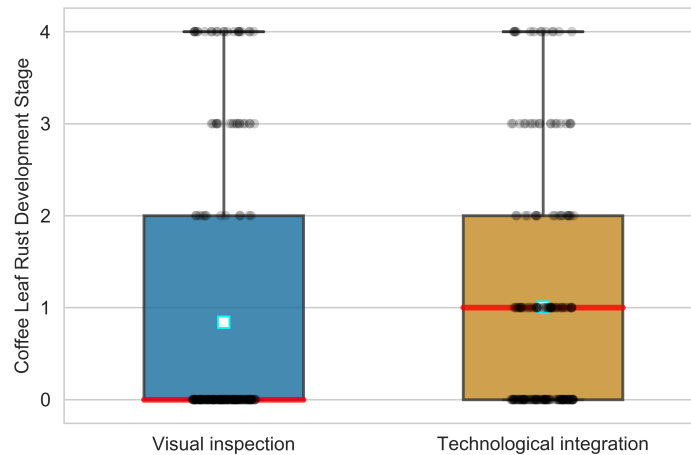
**Table 5.**  $F_1$ -score reached by the individual submodels and the composite model.

Model	$F_1$ -score (Test Set)
Sensor Data (JSON)	0.570
RGB	0.920
RGN	0.946
RE	0.944
<b>Composite</b>	<b>0.775</b>

3.1. Analysis of the Results

Statistical analysis of the results regarding the performance evaluation of the diagnostic model was carried out using the comparative table found in Figure A1 from Appendix A. The purpose of the analysis was to determine whether there was a significant difference in the mean CLR development stage diagnosed with a visual inspection and using the proposed technological integration. The outcome was relevant to get the necessary statistical support for answering the research question.

The comparative table contained 202 observations with the corresponding diagnosed development stage for both treatments. Figure 13 shows the box plot chart describing the measurements. The  $x$ -axis contains the two treatments (“visual inspection” and “technological integration”), whereas the CLR development stage is presented on the  $y$ -axis. The graphical similarity of the data distribution of each treatment suggested a possible similarity to the means of the response variable. To assess this condition and make a decision based on the hypothesis, an ANOVA was executed.



**Figure 13.** Data distribution of the observations for both treatments.

The results of the ANOVA can be seen in Table 6. The obtained  $p$ -value for the treatments factor was 0.231. This value was greater than the set significance ( $\alpha = 0.1$ ), which meant that there was not sufficient evidence for rejecting the null hypothesis. Thus, it was concluded, with 90% confidence, that there was no statistically significant difference between the diagnosis of the CLR development stage made by using visual inspection and the technological integration. This result indicated that both methods were significantly similar to diagnose the disease.

This research demonstrated the feasibility of diagnosing the CLR development stage in the Colombian Caturra variety, with significant performance, through the integration of RS, WSN, and DL. The analysis of the results allowed obtaining statistical evidence for supporting the research hypothesis. In that sense, the outcome suggested that a technological integration could contribute to the protection

of the phytosanitary status of coffee crops since it showed potential for complementing the traditional visual inspections towards the diagnosis of the most economically limiting disease for Colombian coffee production.

**Table 6.** ANOVA table of the statistical analysis.

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Treatments	1	2.7	2.696	1.437	<b>0.231</b>
Residuals	402	753.9	1.875		

#### 4. Conclusions

The integration of RS, WSN, and DL within the framework of this study successfully allowed evaluating to what extent it was possible to diagnose the CLR development stage in the Colombian Caturra variety. To this end, the most relevant information obtained was consolidated, the knowledge about the study context and CLR was detailed, and the repercussions of the disease in the Colombian coffee growing industry were identified. Furthermore, the state-of-the-art methods were reviewed and used for the current research. Creative design sessions were carried out to define the most useful technological integration of RS and WSN. Afterward, a functional prototype that automatically collected data in the field and transferred them to a remote server over the Internet was built. Besides, a diagnostic model using DL was implemented based on the stored data, and it successfully allowed evaluating the CLR development stage with unknown field data.

The motivation of this research project was to contribute to rural development through technological innovation to strengthen the profitability of Colombian coffee growers. Considering that the country has the potential, in terms of environmental conditions and diverse ecosystems, to generate a giant portfolio of exotic products that would be better valued in the specialty coffee market, this research evaluated, with empirical evidence, a technological approach that attempted to facilitate the diagnosis and mitigate the risks of one of the most economically limiting diseases for coffee production. In that sense, the proposed technological integration could positively impact the rural sector since those innovations promote investments in infrastructure, which are crucial to empower the rural community and improve the living standards and activities concerning progress, productivity, and income generation.

The obtained  $p$ -value in the analysis of the results was 0.231, which helped to determine, with 90% confidence, that the visual inspection and the technological integration did not present a statistically significant difference regarding the diagnosis of the CLR development stage. Thus, it could be said that the assessment of the disease led to a similar outcome using either method, which suggested that the obtained results supported the research hypothesis. Finally, it could be asserted that through the integration of RS, WSN, and DL, it was possible to diagnose the CLR development stage in the Colombian Caturra variety with a  $F_1$ -score of 0.775. This value indicated that, on average, the diagnostic model was excellent in terms of the certainty and usefulness of its diagnosis.

Regarding the data processing phase, a further extension of this research could include the implementation of a simple user interface for visualizing the diagnosis of the CLR development stage through the generated DL model to better illustrate the results to a coffee grower. Additionally, the proposed technological integration could be scaled to a real context by using drones with one or both of the two multispectral cameras used in the experiment presented by this work (depending on the project budget) as a possible approach, knowing that the identification of the CLR could be done with just one camera, e.g., RGN ( $F_1$ -score of 0.946), due to its high score. Another real context approach could be further explored using a mobile autonomous robot with a single RGB camera. Finally, the  $F_1$ -score values achieved on the test set, which showed that the submodels based on images presented a higher performance than the JSON submodel (sensor data model), suggested reconsidering the composite model for future work and focusing all efforts on improving the collection and processing of just RGB and multispectral data or using more robust sensors when the technology



allows it; by using just the three submodels (RGB, RGN, and RE), we computed an average  $F_1$ -score of 0.93, which clearly showed that an improved composite  $F_1$ -score could be surely achieved, but a real context commercial application may only implement one of the best three previous submodels due to both implementation and maintenance costs.

**Author Contributions:** A.S. and S.S. designed and implemented the experimental testbed and algorithms for integrating WSN, RS, and DL for CLR detection. D.V. and M.T. supervised the experimental design and managed the project. M.M. and B.S. reviewed the machine learning and deep learning parts of this research. All authors contributed to the writing and reviewing of the present manuscript. All authors read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Universidad EAFIT internal research projects and Colciencias “Jovenes Investigadores e Innovadores por la paz 2017”.

**Acknowledgments:** We would like to thank Engineer Carlos Mario Ospina for his support during the visit to CENICAFÉ’s Experimental Station “El Rosario”, as well as coffee grower Nabor Giraldo, Diego Miguel Sierra, and Dentist Samuel Roldán for their support in the procurement of the coffee plants and the biological matter containing CLR. Additionally, we would like to thank Alejandro Marulanda, Edwin Nelson Montoya, Engineer Hugo Murillo, Universidad EAFIT, and Colciencias for the constant academic, infrastructural, and financial support. It is also important to highlight the enormous work and dedication of Project Supervisors Engineer Camilo Velásquez, Engineer Felipe Gutiérrez, Engineer Sebastián Osorio, Engineer Juan Diego Zuluaga, and physics engineering undergraduate student Sergio Jurado. In addition, we want to acknowledge the biology team, formed of undergraduate students Alisson Martínez and Laura Cristina Moreno and led by Luisa Fernanda Posada. We would also like to mention computing-systems engineering undergraduate students Alejandro Cano, Luis Javier Palacio, and Sebastián Giraldo, and we also thank the Engineers Ricardo Urrego and Luis Felipe Machado and Sebastián Rodríguez.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

CLR	Coffee Leaf Rust
FCP	Fondo Colombia en Paz
RS	Remote Sensing
WSN	Wireless Sensor Networks
ML	Machine Learning
DL	Deep Learning
LAI	Leaf Area Index
CASI	Compact Airborne Spectrographic Imager
NIR	Near-Infrared
RGB	Red Green Blue
RE	Red Edge
RGN	Red Green Near-Infrared
NDVI	Normalized Difference Vegetation Index
UAV	Unmanned Aerial Vehicle
BS	Base Station
PSoC	Programmable System on a Chip
AI	Artificial Intelligence
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
MSI	Multispectral Instrument
CRD	Completely Randomized Design
ANOVA	Analysis Of Variance
SBC	Single Board Computer
SFTP	Secure File Transfer Protocol
VNC	Virtual Network Computing
SSH	Secure Shell
IoT	Internet of Things
JSON	JavaScript Object Notation



Appendix B. Data Management Model

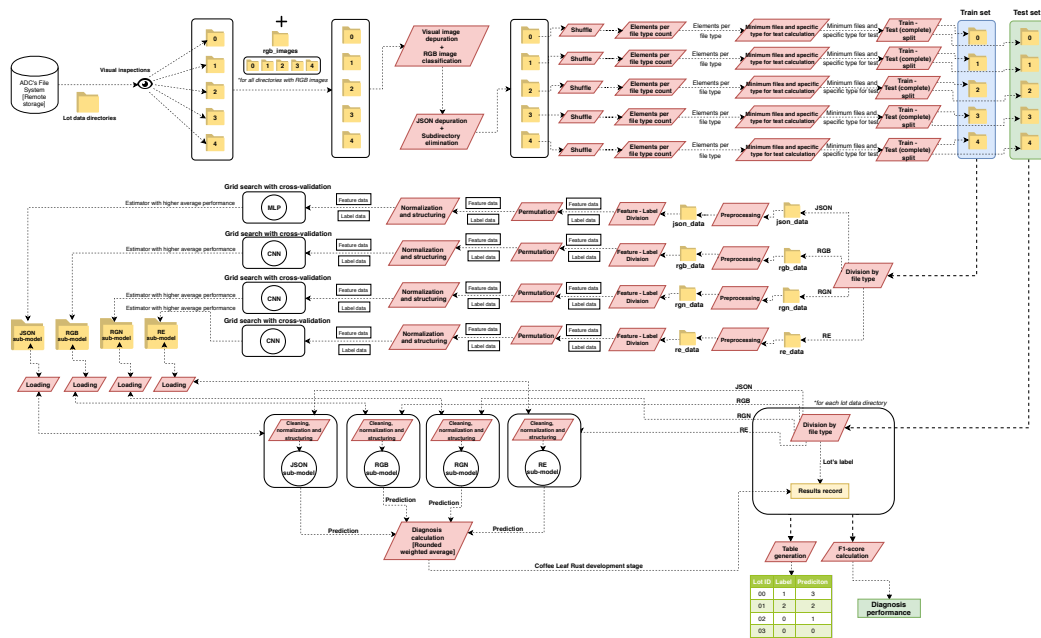


Figure A2. Detailed Data Management Model.

Appendix C. IoT Platform Dashboard



Figure A3. Implemented IoT platform real-time dashboard.

Appendix D. Submodels' Architectures

Table A1. JSON submodel's architecture.

N	Layer	Output Shape	# of Parameters
1.	Input Layer	(None, 6)	-
2.	Fully Connected	(None, 16)	112
3.	Batch Normalization	(None, 16)	64
4.	Activation	(None, 16)	0
5.	Fully Connected	(None, 64)	1088
6.	Batch Normalization	(None, 64)	256
7.	Activation	(None, 64)	0
8.	Dropout	(None, 64)	0
9.	Fully Connected	(None, 32)	2080
10.	Batch Normalization	(None, 32)	128
11.	Activation	(None, 32)	0
12.	Dropout (rate = rate/2)	(None, 32)	0
13.	Fully Connected	(None, 4)	132
14.	Activation	(None, 4)	0

**Table A2.** RGB submodel’s architecture.

N	Layer	Output Shape	# of Parameters
1.	Input Layer	(None, 96, 128, 3)	-
2.	Convolutional2D (Kernel = (5, 5))	(None, 92, 124, 18)	1368
3.	Batch Normalization	(None, 92, 124, 18)	72
4.	Activation	(None, 92, 124, 18)	0
5.	Max Pooling (pool = (2, 2))	(None, 46, 62, 18)	0
6.	Convolutional2D (kernel = (5, 5))	(None, 42, 58, 36)	16,236
7.	Batch Normalization	(None, 42, 58, 36)	144
8.	Activation	(None, 42, 58, 36)	0
9.	Max Pooling (pool = (2, 2))	(None, 21, 29, 36)	0
10.	Convolutional2D (kernel = (3, 3))	(None, 19, 27, 54)	17,550
11.	Batch Normalization	(None, 19, 27, 54)	216
12.	Activation	(None, 19, 27, 54)	0
13.	Max Pooling (pool = (2, 2))	(None, 9, 13, 54)	0
14.	Dropout	(None, 9, 13, 54)	0
15.	Flatten	(None, 6318)	0
16.	Fully Connected	(None, 512)	3,235,328
17.	Batch Normalization	(None, 512)	2048
18.	Activation	(None, 512)	0
19.	Dropout	(None, 512)	0
20.	Fully Connected	(None, 128)	65,664
21.	Batch Normalization	(None, 128)	512
22.	Activation	(None, 128)	0
23.	Dropout (rate = rate/2)	(None, 128)	0
24.	Fully Connected	(None, 5)	645
25.	Activation	(None, 5)	0

**Table A3.** RGN and RE submodels’ architectures.

N	Layers	Output Shape	# of Parameters
1.	Input Layer	(None, 128, 96, 3)	-
2.	Convolutional2D (kernel = (5, 5))	(None, 124, 92, 18)	1368
3.	Batch Normalization	(None, 124, 92, 18)	72
4.	Activation	(None, 124, 92, 18)	0
5.	Max Pooling (pool = (2, 2))	(None, 62, 46, 18)	0
6.	Convolutional2D (kernel = (5, 5))	(None, 58, 42, 36)	16,236
7.	Batch Normalization	(None, 58, 42, 36)	144
8.	Activation	(None, 58, 42, 36)	0
9.	Max Pooling (pool = (2, 2))	(None, 29, 21, 36)	0
10.	Convolutional2D (kernel = (3, 3))	(None, 27, 19, 54)	17,550
11.	Batch Normalization	(None, 27, 19, 54)	216
12.	Activation	(None, 27, 19, 54)	0
13.	Max Pooling (pool = (2, 2))	(None, 13, 9, 54)	0
14.	Dropout	(None, 13, 9, 54)	0
15.	Flatten	(None, 6318)	0
16.	Fully Connected	(None, 512)	3,235,328
17.	Batch Normalization	(None, 512)	2048
18.	Activation	(None, 512)	0
19.	Dropout	(None, 512)	0
20.	Fully Connected	(None, 128)	65,664
21.	Batch Normalization	(None, 128)	512
22.	Activation	(None, 128)	0
23.	Dropout (rate = rate/2)	(None, 128)	0
24.	Fully Connected	(None, 4)	516
25.	Activation	(None, 4)	0

## References

- Esposito, M. *Driving the Sustainability of Production Systems with Fourth Industrial Revolution Innovation*; White Paper; World Economic Forum: Colony, Switzerland, 2018.
- OECD Directorate for Science, Technology and Industry. *ISIC REV. 3 TECHNOLOGY INTENSITY DEFINITION: Classification of Manufacturing Industries into Categories Based on R&D Intensities*; OECD: Paris, France, 2011.
- Chen, Y.; Zhang, C.; Wang, S.; Li, J.; Li, F.; Yang, X.; Wang, Y.; Yin, L. Extracting Crop Spatial Distribution from Gaofen 2 Imagery Using a Convolutional Neural Network. *Appl. Sci.* **2019**, *9*, 2917. [[CrossRef](#)]
- Liu, J.; Zhang, X.; Li, Z.; Zhang, X.; Jemric, T.; Wang, X. Quality Monitoring and Analysis of Xinjiang 'Korla' Fragrant Pear in Cold Chain Logistics and Home Storage with Multi-Sensor Technology. *Appl. Sci.* **2019**, *9*, 3895. [[CrossRef](#)]
- Lee, J.W.; Kim, S.C.; Oh, J.; Chung, W.J.; Han, H.W.; Kim, J.T.; Park, Y.J. Engine Speed Control System for Improving the Fuel Efficiency of Agricultural Tractors for Plowing Operations. *Appl. Sci.* **2019**, *9*, 3898. [[CrossRef](#)]
- Zhou, C.; Ye, H.; Xu, Z.; Hu, J.; Shi, X.; Hua, S.; Yue, J.; Yang, G. Estimating Maize-Leaf Coverage in Field Conditions by Applying a Machine Learning Algorithm to UAV Remote Sensing Images. *Appl. Sci.* **2019**, *9*, 2389. [[CrossRef](#)]
- National Coffee Association USA. The Influence of Coffee Around the World. 2015. Available online: <https://nationalcoffeeblog.org/2015/06/15/the-influence-of-coffee-around-the-world/> (accessed on 8 May 2018).
- SCAA. *SCAA Protocols Cupping Specialty Coffee*; Specialty Coffee Association of America: Santa Ana, CA, USA, 2015.
- CropLife Latin America. *Roya del cafeto*. 2018. Available online: <https://www.croplifela.org/es/plagas/listado-de-plagas/roya-del-cafeto> (accessed on 19 January 2019).
- Rivillas, C.; Serna, C.; Cristancho, M.; Gaitan, A. *La Roya del Cafeto en Colombia: Impacto Manejo y Costos del Control*; Avances Tecnicos Cenicafe: Chinchiná, Colombia, 2011.
- Carvalho, C.R.; Fernandes, R.C.; Carvalho, G.M.A.; Barreto, R.W.; Evans, H.C. Cryptosexuality and the Genetic Diversity Paradox in Coffee Rust, *Hemileia vastatrix*. *PLoS ONE* **2011**. [[CrossRef](#)]
- The Observatory of Economic Complexity (OEC). Colombia (COL) Exports, Imports, and Trade Partners. 2018. Available online: <https://atlas.media.mit.edu/en/profile/country/col/> (accessed on 23 January 2019).
- Federación Nacional de Cafeteros. *Estadísticas Historicas*. 2018. Available online: [https://www.federaciondecafeteros.org/particulares/es/quienes\\_somos/119\\_estadisticas\\_historicas/](https://www.federaciondecafeteros.org/particulares/es/quienes_somos/119_estadisticas_historicas/) (accessed on 10 January 2019).
- Guzmán, O.; Gómez, E.; Rivillas, C.; Oliveros, C. Utilización del procesamiento de imágenes para determinar la severidad de La Mancha de Hierro, en hojas de café. *Cenicafé* **2003**, *54*, 258–265.
- Martinelli, F.; Scalenghe, R.; Davino, S.; Panno, S.; Scuderi, G.; Ruisi, P.; Villa, P.; Stroppiana, D.; Boschetti, M.; Goulart, L.R.; et al. Advanced methods of plant disease detection. A review. *Agron. Sustain. Dev.* **2015**, *35*, 1–25. [[CrossRef](#)]
- Lobitz, B.; Beck, L.; Huq, A.; Wood, B.; Fuchs, G.; Faruque, A.; Colwell, R. Climate and infectious disease: Use of remote sensing for detection of *Vibrio cholerae* by indirect measurement. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 1438–1443. [[CrossRef](#)]
- Su, N.Y. Remote Monitoring System for Detecting Termites. U.S. Patent 6,052,066, 29 September 1998.
- Mirik, M.; Norland, J.E.; Crabtree, R.L.; Biondini, M.E. Hyperspectral one-meter-resolution remote sensing in Yellowstone National Park, Wyoming: I. Forage nutritional values. *Rangel. Ecol. Manag.* **2005**, *58*, 452–458. [[CrossRef](#)]
- Mulla, D.J. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosyst. Eng.* **2013**, *114*, 358–371, doi:10.1016/j.biosystemseng.2012.08.009. [[CrossRef](#)]
- Calvario, G.; Sierra, B.; Alarcón, T.E.; Hernandez, C.; Dalmau, O. A multi-disciplinary approach to remote sensing through low-cost UAVs. *Sensors* **2017**. [[CrossRef](#)]
- Goel, P.K.; Prasher, S.O.; Landry, J.A.; Patel, R.M.; Bonnell, R.; Viau, A.A.; Miller, J. Potential of airborne hyperspectral remote sensing to detect nitrogen deficiency and weed infestation in corn. *Comput. Electron. Agric.* **2003**, *38*, 99–124. [[CrossRef](#)]

22. Ortega-Huerta, M.A.; Komar, O.; Price, K.P.; Ventura, H.J. Mapping coffee plantations with Landsat imagery: An example from El Salvador. *Int. J. Remote Sens.* **2012**, *33*, 220–242. [[CrossRef](#)]
23. Bolaños, J.A.; Campo, L.; Corrales, J.C. Characterization in the Visible and Infrared Spectrum of Agricultural Crops from a Multicopter Air Vehicle. In Proceedings of the International Conference of ICT for Adapting Agriculture to Climate Change, Popayán, Colombia, 22–24 November 2017; pp. 29–43.
24. Piedallu, C.; Cheret, V.; Denux, J.; Perez, V.; Azcona, J.; Seynave, I.; Gégout, J. *Etudier les Variations Spatiales de NDVI pour Caractériser les Contraintes Environnementales Limitant la Vitalité des Forêts de Montagne et de Méditerranée*; CAQSI; INRA: Clermont-Ferrand, France, 2018; pp. 1–17.
25. Chemura, A.; Mutanga, O.; Dube, T. Remote sensing leaf water stress in coffee (*Coffea arabica*) using secondary effects of water absorption and random forests. *Phys. Chem. Earth Parts A/B/C* **2017**, *100*, 317–324. [[CrossRef](#)]
26. Hamuda, E.; Glavin, M.; Jones, E. A survey of image processing techniques for plant extraction and segmentation in the field. *Comput. Electron. Agric.* **2016**, *125*, 184–199. [[CrossRef](#)]
27. Camargo, A.; Smith, J.S. An image-processing based algorithm to automatically identify plant disease visual symptoms. *Biosyst. Eng.* **2009**. [[CrossRef](#)]
28. De Melo Virginio Filho, E.; Astorga, C. *Prevención y Control de la Roya del Café: Manual de Buenas Prácticas para Técnicos y Facilitadores*, 1st ed.; CATIE: Turrialba, Costa Rica, 2015; p. 67.
29. Avelino, J.; Zelaya, H.; Merlo, A.; Pineda, A.; Ordoñez, M.; Savary, S. The intensity of a coffee rust epidemic is dependent on production situations. *Ecol. Model.* **2006**, *197*, 431–447. [[CrossRef](#)]
30. Haddad, F.; Maffia, L.A.; Mizubuti, E.S.; Teixeira, H. Biological control of coffee rust by antagonistic bacteria under field conditions in Brazil. *Biol. Control* **2009**, *49*, 114–119. [[CrossRef](#)]
31. Jackson, D.; Skillman, J.; Vandermeer, J. Indirect biological control of the coffee leaf rust, *Hemileia vastatrix*, by the entomogenous fungus *Lecanicillium lecanii* in a complex coffee agroecosystem. *Biol. Control* **2012**, *61*, 89–97. [[CrossRef](#)]
32. Azfar, S.; Nadeem, A.; Basit, A. Pest detection and control techniques using wireless sensor network: A review. *J. Entomol. Zool. Stud.* **2015**, *3*, 92–99.
33. Dargie, W.; Poellabauer, C. *Fundamentals of Wireless Sensor Networks: Theory and Practice*; Wireless Communications and Mobile Computing; Wiley: Hoboken, NJ, USA, 2010.
34. Chaudhary, D.; Nayse, S.; Waghmare, L. Application of wireless sensor networks for greenhouse parameter control in precision agriculture. *Int. J. Wirel. Mob. Networks (IJWMN)* **2011**, *3*, 140–149. [[CrossRef](#)]
35. Piamonte, M.; Huerta, M.; Clotet, R.; Padilla, J.; Vargas, T.; Rivas, D. WSN Prototype for African Oil Palm Bud Rot Monitoring. In Proceedings of the International Conference of ICT for Adapting Agriculture to Climate Change, Popayán, Colombia, 22–24 November 2017; pp. 170–181.
36. Bhardwaj, A.; Di, W.; Wei, J. *Deep Learning Essentials: Your Hands-On Guide to the Fundamentals of Deep Learning and Neural Network Modeling*; Packt Publishing: Birmingham, UK, 2018.
37. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; Adaptive Computation and Machine Learning; MIT Press: Cambridge, MA, USA, 2016.
38. Patterson, J.; Gibson, A. *Deep Learning: A Practitioner's Approach*; O'Reilly Media: Sebastopol, CA, USA, 2017.
39. Sulisty, S.B.; Wu, D.; Woo, W.L.; Dlay, S.S.; Gao, B. Computational Deep Intelligence Vision Sensing for Nutrient Content Estimation in Agricultural Automation. *IEEE Trans. Autom. Sci. Eng.* **2018**. [[CrossRef](#)]
40. Sulisty, S.B.; Woo, W.L.; Dlay, S.S. Regularized Neural Networks Fusion and Genetic Algorithm Based On-Field Nitrogen Status Estimation of Wheat Plants. *IEEE Trans. Ind. Inform.* **2017**. [[CrossRef](#)]
41. Sulisty, S.B.; Woo, W.L.; Dlay, S.S.; Gao, B. Building a Globally Optimized Computational Intelligent Image Processing Algorithm for On-Site Inference of Nitrogen in Plants. *IEEE Intell. Syst.* **2018**. [[CrossRef](#)]
42. Fuentes, A.; Yoon, S.; Kim, S.C.; Park, D.S. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors* **2017**. [[CrossRef](#)]
43. Picon, A.; Alvarez-Gila, A.; Seitz, M.; Ortiz-Barredo, A.; Echazarra, J.; Johannes, A. Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. *Comput. Electron. Agric.* **2019**. [[CrossRef](#)]
44. Chemura, A.; Mutanga, O.; Dube, T. Separability of coffee leaf rust infection levels with machine learning methods at Sentinel-2 MSI spectral resolutions. *Precis. Agric.* **2017**, *18*, 859–881. [[CrossRef](#)]
45. Chemura, A.; Mutanga, O.; Sibanda, M.; Chidoko, P. Machine learning prediction of coffee rust severity on leaves using spectroradiometer data. *Trop. Plant Pathol.* **2018**, *43*, 117–127. [[CrossRef](#)]

46. Mollazade, K.; Omid, M.; Tab, F.A.; Mohtasebi, S.S. Principles and Applications of Light Backscattering Imaging in Quality Evaluation of Agro-food Products: A Review. *Food Bioprocess Technol.* **2012**, *5*, 1465–1485. [[CrossRef](#)]
47. Pulido, H.G.; De la Vara Salazar, R.; González, P.G.; Martínez, C.T.; Pérez, M.d.C.T. *Análisis y Diseño de Experimentos*; McGraw-Hill: New York, NY, USA, 2012.
48. Numpy.org. NumPy. 2018. Available online: <http://www.numpy.org/> (accessed on 19 January 2019).
49. Pandas.pydata.org. Pandas: Powerful Python Data Analysis Toolkit. 2018. Available online: <https://pandas.pydata.org/> (accessed on 19 January 2019).
50. Scikit-learn.org. Scikit-learn. 2018. Available online: <https://scikit-learn.org/stable/> (accessed on 19 January 2019).
51. Keras.io. Keras: The Python Deep Learning library. 2018. Available online: <https://keras.io/> (accessed on 19 January 2019).
52. Tensorflow.org. TensorFlow. 2018. Available online: <https://www.tensorflow.org/> (accessed on 19 January 2019).
53. Burkov, A. *The Hundred-Page Machine Learning Book*; Andriy Burkov: Quebec, QC, Canada, 2019.
54. scikit-learn Developers. `sklearn.model_selection.GridSearchCV`. 2019. Available online: [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.GridSearchCV.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html) (accessed on 19 January 2019).
55. scikit-learn Developers. 3.1. Cross-Validation: Evaluating Estimator Performance. 2019. Available online: [https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html) (accessed on 19 January 2019).



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).



Received 9 June 2022, accepted 29 June 2022, date of publication 4 July 2022, date of current version 13 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3188102

 RESEARCH ARTICLE

# A Hybrid Machine-Learning Ensemble for Anomaly Detection in Real-Time Industry 4.0 Systems

DAVID VELÁSQUEZ<sup>1,2,3</sup>, ENRIQUE PÉREZ<sup>2</sup>, XABIER OREGUI<sup>2</sup>, ARKAITZ ARTETXE<sup>2</sup>, JORGE MANTECA<sup>4</sup>, JORDI ESCAYOLA MANSILLA<sup>5</sup>, MAURICIO TORO<sup>1</sup>, MIKEL MAIZA<sup>2</sup>, AND BASILIO SIERRA<sup>1,3</sup>

<sup>1</sup>RID on Information Technologies and Communications Research Group, Universidad EAFIT, Medellín 050022, Colombia

<sup>2</sup>Department of Data Intelligence for Energy and Industrial Processes, Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), 20014 Donostia-San Sebastián, Spain

<sup>3</sup>Department of Computer Science and Artificial Intelligence, University of Basque Country (UPV/EHU), 20018 Donostia-San Sebastián, Spain

<sup>4</sup>Technical Department Direction, Mapner, 20115 Astigarraga, Spain

<sup>5</sup>Department of Statistics and Operational Research, Universitat Oberta de Catalunya, Rambla del Poblenou, 08018 Barcelona, Spain

Corresponding author: David Velásquez (dvelas25@eafit.edu.co)

This work was supported in part by Vicomtech Foundation and in part by Universidad EAFIT.

**ABSTRACT** Detecting faults and anomalies in real-time industrial systems is a challenge due to the difficulty of sufficiently covering an industrial system's complexity. Today, Industry 4.0 makes it possible to tackle these problems through emerging technologies such as the Internet of Things and Machine Learning. This paper proposes a hybrid machine-learning ensemble real-time anomaly-detection pipeline that combines three Machine Learning models –Local Outlier Factor, One-Class Support Vector Machine, and Autoencoder–, through a weighted average to improve anomaly detection. The ensemble model was tested with three air-blowing machines obtaining a  $F_1$ -score value of 0.904, 0.890, and 0.887, respectively. The results of the ensemble model showed improved performance metrics concerning the individual metrics. A novelty of this model is that it consists of two stages inspired by a standard industrial system: i) a manufacturing stage and ii) an operation stage.

**INDEX TERMS** Anomaly detection, industry 4.0, machine learning, predictive maintenance, real-time.

## I. INTRODUCTION

Thanks to the fourth industrial revolution (4IR), traditional industrial processes face new challenges: improving current or establishing new processes that efficiently use novel technologies and fully exploit their potential. 4IR or Industry 4.0 is viewed as a disruptive innovation in a highly competitive market that positively impacts several industrial sectors by incorporating new enabling technologies: 3D printing, the Internet of Things (IoT), Cyber-Physical Systems (CPS), Artificial Intelligence (AI), Big Data, Robotics, Nanotechnology, and Quantum Computing are examples of these technologies [1]. In industrial machines, high volumes of data are

generated and acquired by data acquisition systems such as a Supervisory Control and Data Acquisition (SCADA) or an embedded system. AI algorithms can then process this data to generate new knowledge of the process and identify new machine conditions, which represents one of the advancements provided by Industry 4.0. Predictive maintenance is an industrial process that is the subject of the work presented in this article and highly benefits from the Industry 4.0 technologies mentioned above [2].

Nowadays, most industrial companies face problems arising from maintaining their systems. However, multiple techniques –involving predictive or *condition-based maintenance (CBM)*– allow predicting critical situations to reduce these problems. According to An *et al.* [3], in terms of diagnosis, predictive maintenance is divided into two categories:

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval<sup>1</sup>.

i) Models that take into account physical principles and ii) models based on historical observations. One of the techniques used in the second group consists of the early detection of abnormal behavior in industrial equipment. This early detection can avoid possible breakdowns of equipment and reduce associated maintenance costs.

Anomaly detection is being researched in several application fields. Some of the associated research fields are disease detection, intrusion detection, fraud prediction, and fault detection in industrial equipment [4]. It is possible to identify anomalous states that do not match the normality data, which usually corresponds to the predominant states through anomaly detection.

The detection of anomalous states presents a challenging task. The detection becomes more complicated than usual if it is to be done in real-time due to the restrictive features of the streaming data. Unlike batch learning, where all the historical data are available, and no new information is added to the models already built, stream learning has five restrictions that must be taken into account [5]. i) Streaming data samples arrive online and can be read at most one time, which is a strong restriction for processing them since the system has to decide whether the current data sample is discarded or archived. ii) Past data samples can only be accessed if stored in memory. Otherwise, a forgetting mechanism in charge of discarding past samples is applied. iii) Since not all data samples can be stored, a decision made on past samples cannot be undone. iv) The data processing time of each data sample should be short and constant. v) The data processing algorithm must produce a model equivalent to what a batch algorithm would produce.

The former five restrictions are why most anomaly detection algorithms—for batch processing—do not apply to stream processing. Nonetheless, there are hybrid approaches that use batch-learning algorithms to build an initial model as the first step and then apply streaming anomaly-detection algorithms as the second step.

The contribution of this work is the evaluation and comparison of different methods to detect anomalies that, due to their performance-control metrics, establish the weight (or incidence) of each method in the final combined model, thus responding better and efficiently to the challenge of real-time anomaly detection. Specifically, the present work combines the predicted output of three Machine Learning (ML) models: Local Outlier Factor (LOF), One-Class Support Vector Machine (OCSVM), and Autoencoder employing a weighted average—using as weight the  $F_1$ -score value of each model. The goal of the combined model is the detection of anomalies in industrial systems in real-time. The proposed hybrid model was implemented using a data set from a real industrial system of air-blowing machines. Thus, it can be said that the proposed hybrid anomaly detection model applies to Industry 4.0 systems as well as other industrial frameworks where real-time data acquisition systems are available.

The following sections of the article are divided into four sections. The state-of-the-art section shows existing

approaches and research for anomaly detection in real-time. Next, the third section shows a detailed explanation of the proposed hybrid anomaly detection. Finally, the results section describes the scores obtained by applying the hybrid anomaly detection methodology to a testing data set. A Conclusions section ends this paper, showing some concluding remarks and a future work proposal.

## II. STATE OF THE ART

According to [6], [7], an anomaly can be defined as a point in time where the system's behavior is unusual and significantly different from previous, normal behavior. An anomaly may imply an adverse change in the system, for instance, a fluctuation in a jet engine's turbine rotation frequency, which possibly means an imminent failure. An anomaly may also mean positive behavior; for instance, many web clicks on a new product page imply higher demand. In both cases, anomalies in data provide an insight into abnormal behavior that can be translated into potentially useful information.

The challenge of detecting anomalies—in an industrial environment—can be twofold. Firstly, to propose a method to understand different data obtained from various sensors, often with excessive noise. Secondly, to obtain an overview of normal behavior to characterize such behavior from historical data. Therefore, to correctly detect anomalies in a data set, one must first characterize and define normal data behavior [8]. In addition, normal behavior can be characterized by the following three stages. (i) Consider data describing normal behavior through historical data (without considering anomalies) segmented into different classes according to the context in which they were recorded. (ii) Extract the most frequent behaviors, thus characterizing each class. (iii) Detect anomalies in newly recorded data based on previous knowledge.

In general, anomalies are classified into three types: specific, contextual, and collective [9]–[11]. It is considered a point anomaly when this single data point is recognized as anomalous concerning the rest of the data. According to [10], these anomalies must be identified before processing or analyzing the data.

- *Contextual anomalies* are those where the data are considered anomalous in a specific context (e.g., the same sample data are “normal” in a given scenario but anomalous in another context). These types of anomalies are more common in time-series data flows [10].
- *Collective anomalies* are those that occur when a collection of related data are considered anomalous to the total data. Collective anomalies can also be spatial if they are outside a typical range or temporal, where the value is not outside the typical range. However, the sequence in which it occurs is unusual.

Anomaly detection methods can be distinguished as supervised, semi-supervised and unsupervised. Using one method or another usually depends on the existence or not of descriptive labels of the anomaly. The labels can be categorical,

e.g., we can have a case of binary or all/nothing labels such as “anomalous behavior” (1) and “non-anomalous / normal behaviour (0)”, or numerical, e.g., a value of “anomaly score” ranging from 0 (“non-anomalous / normal”) to 1 (“totally anomalous”). While anomaly detection could be posed as a supervised learning problem, this is –generally– not the case, as there is often no or little data labeled with the anomalous behavior [12].

Once the data is available, normally, a series of transformations of the data needs to be performed before starting the anomaly detection process [13].

- *Aggregation methods*: A set of consecutive values from a time-series data is replaced by a corresponding representative value. It provides benefits such as reducing dimensionality, although it can make detecting anomalies in subsequent steps difficult.
- *Discretisation methods*: Time-series data are converted into a discrete sequence of finite alphabets. Techniques such as symbolic sequence and editing distance can be applied to detect anomalies.
- *Digital Signal Processing (DSP) techniques* (such as Fourier transform, Gabor, and Wavelets filters): Time-series data are transformed into a lower-dimensional representation of the input data where anomaly detection can take place.

A common type of problem detected, which may be present in the data, is noise and outliers. Noise among normal data may cause the model not to obtain the desired optimal predictions. Outliers are data points that may be caused by noise or may have an irregular pattern of behavior. Therefore, this unusual behavior must first be identified and decided whether it should be considered an anomaly or an outlier.

Usually, data are created by one or more generation processes, representing system’s activities. When the generation process behaves unusually, it creates anomalies. Therefore, an anomaly often contains valuable information about the abnormal characteristics of the systems and elements that impact the generation process [11].

### A. CLASSIFICATION OF TECHNIQUES FOR ANOMALY DETECTION

There are currently six techniques to detect anomalies. These techniques are i) Statistics, ii) Classification, iii) Clustering, iv) Similarity-based, v) Soft Computing, and vi) Knowledge and Combined Techniques based, as explained in [13]. In Table 1, these techniques –and some examples of the algorithms– used can be seen in detail. The most relevant ones for this work will be detailed next.

#### 1) STATISTICS BASED ANOMALY DETECTION TECHNIQUES

Statistical techniques adjust a predefined distribution to a given data and apply statistical inference to determine whether an instance belongs to that model. Instances with a low probability are reported as anomalies [14].

**TABLE 1. Classification of the different techniques for anomaly detection [13].**

Technique	Sub Techniques	Examples
Statistical	Parametric, Non-parametric	Box-plot, Grubbs test, Chi-square, PCA, Kernel methods
Classification	Multi-Class, One Class	Neural Networks, Bayesian Networks, SVM, Decision Trees
Clustering	Parametric, Non-parametric	DBSCAN, Rock, SNN, K-means, EM, LOF variants
Similarity based	Continuous and categorical data	k-NN variants, Relative Density
Soft Computing	GA, NN, Fuzzy and Rough Sets, Ant Colony	GANIDS, NN, DNN, CNN
Knowledge based	Rules and Expert Systems, Ontology and Logic-based techniques	Decision Trees
Combination Learners	Ensemble based, Fusion based, Hybrid	Bagging and Boosting

The two typologies covered by this technique are parametric and non-parametric. The first assumes an underlying data distribution. Although somewhat less efficient in finding anomalies, the second is preferred because, a priori, it does not define any model structure as this is determined from the data.

The most common parametric techniques are divided into those based on Gaussian models and those based on regression models. If a non-parametric approach is to be followed, such a classification can be made based on histograms or kernels.

Statistical techniques work well for simple structured data with small dimensions and volume. In such cases, several methods can be used [13], such as Box-plots, Blum Floyd Pratt Rivest Tarjan (BFPR) algorithm, and similar central-value estimations on data streams; Medcouple and Grubbs test (for univariate data); Comparison of distributions (QQ charts, Kolmogorov-Smirnov test, Kruskal-Wallis test, and Wilcoxon signed range tests); Auto-regressive techniques (Auto-regressive Integrated Moving Average - ARIMA, Auto-regressive Moving Average - ARMA); ML-based methods; Bayesian networks. Principal Components Analysis (PCA) / Independent Component Analysis (ICA) (e.g., sequence micro-batch analysis).

#### 2) CLASSIFICATION BASED ANOMALY DETECTION TECHNIQUES

Classification-based anomaly detection techniques perform two main stages called training and testing. *In the training phase*, the system learns from the available samples and generates a classifier. *In the testing phase*, samples that the classifier has not seen are tested to measure the model’s performance. According to the labels available for training, classifiers can be grouped into two categories: i) one-class

and ii) multi-class. Examples of single and multi-class classifiers are neural networks, Bayesian networks, Support Vector Machines (SVM), and decision trees. These, together with fuzzy logic, are also methods that present a good performance in the presence of strong noise [15]–[18].

Classification-based techniques have the advantage of being able to distinguish between observations that belong to different anomalies (instead of an overall class called “anomaly”), and their testing phase is quick, as the test instance is compared to the predefined model [19]. Although, classification techniques are based on the availability of assigning labels to various normal and abnormal classes, which is a difficult task. Also, these techniques assign labels to test data, which can be a disadvantage when an anomaly score is desired.

Classification-based techniques can also be categorized according to the type of anomaly. Radial-Base Functions (RBF), SVM, and derivatives are commonly used for individual anomalies. RBFs are very accurate and fast, particularly for the supervised classification of individual anomalies. For multiple anomalies, Deep Neural Networks (DNN), induction rules, and decision trees are used. DNNs can provide exceptional recognition rates in static scenarios but can give data problems that vary over time.

### 3) CLUSTERING-BASED ANOMALY DETECTION TECHNIQUES

Clustering techniques are generally divided into two stages: first, the data are grouped with clustering algorithms, and then the degree of deviation is analyzed according to the results obtained by the clustering [4]. There are some prior considerations about the data instances in these unsupervised techniques. On the one hand, normal-data samples belong to global clusters. On the other hand, anomalies do not belong to any defined cluster. In addition, normal data samples are near the centroids of the closest cluster, while anomalous data are further away. Finally, normal-data samples belong to large, dense groups, but anomalies belong to local, small, disparate groups.

Cluster-based methods are applied in both supervised and unsupervised learning. Most techniques work well for complex, large-sized, and voluminous data and –optimally– if the anomalies do not form significant clusters in a short time series. Examples of this type of algorithm are k-Means, Shared Nearest Neighbour (SNN), Density-Based Spatial Clustering of Applications with Noise (DBScan), Self-Organizing Map (SOM), or Clustering-based Dynamic indexing Tree (CD-Tree) [4].

### 4) SIMILARITY BASED ANOMALY DETECTION TECHNIQUES

These techniques are the most widely used to detect anomalies. One of the techniques, based on similarity, is known as k Nearest Neighbours (k-NN). k-NN is a non-parametric method that requires a distance metric to measure the similarity between data observations. Although Euclidean distance is the most commonly used metric for data with continuous attributes, it is not usually employed on a practical level.

The above is because the Euclidean distance does not work well in high-dimensional sets, and measurements such as Mahalanobis, Hamming, or Chebyshev distances are used instead. The k-NN algorithm is based on the data score given by the distance to most of the data around it. So, new data are classified according to this score. Although, there are some considerations to be taken into account in this type of technique [13]: i) A shortage of data can be seen as an anomaly in unsupervised techniques. ii) The performance is a function of the distance method chosen; therefore, the criteria must be clear when choosing a metric. iii) It is valid only in cases of low-dimensional data. Defining a measure of the distance between instances can be complicated when the data dimension is increased.

Another essential similarity-based anomaly detection technique is based on relative density rather than distance. This technique estimates the neighborhoods’ density so that a data item in a low-density neighborhood will be anomalous while one in a high-density neighborhood will be considered normal. An existing method for the above is the Local-Outlier Factor (LOF), which introduces the concept of local outliers and is based on scoring a data sample according to the average ratio of the neighborhood’s density to the instance’s density [20].

### B. RELATED WORKS

Many studies on anomaly detection in static data sets in the literature exist. Examples of supervised approaches are SVM and Decision Tree [12], or cluster-based methods such as the Distributed Matching-based Grouping Algorithm (DMGA) [21]. Other examples use self-adaptive and dynamic clustering to learn weights for anomaly detection [22] or statistical methods such as auto-regressive techniques (e.g., ARIMA models [23]).

The problem with these methods is that they are not designed to process streaming data as they need to have the data set previously stored in the main memory. Therefore, these traditional techniques have been adapted first and then applied to streaming-data environments in many cases.

In this sense, Tan *et al.* [24] propose a fast-anomaly detection of a class that uses only normal data and works well when anomalous data are rare. To do this, they use the Half-Space Trees (HS-Trees) algorithm. The HS-Trees algorithm presents a set of random HS trees. Each HS tree consists of a set of nodes, where each node captures the number of data elements (called mass) within a subspace of the data stream. The mass is used to profile the degree of an anomaly as it is quick and straightforward to calculate compared to other methods based on distance or density. The tree structure is constructed without any data, making it very efficient as it does not require restructuring the model once it is running on streaming data. HS-Trees only need normal data for training.

Another technique that is worth mentioning is the isolation-Forest Algorithm for Streaming Data (iForestASD) [25], based on the Isolation-Forest algorithm [26]. This method handles streaming data using sliding

windows. In this case, the authors start from the “concept drift”, which is a common occurrence handling the streaming of data in dynamic and non-stationary environments producing a change in the distribution of the data [27]. The “concept drift” is a problem that occurs when the statistical properties of the target variable change over time and the anomaly detection model is no longer compatible with the data the model handles, resulting in less accurate predictions. Therefore, to maintain the anomaly detection effectively, the model needs to be retrained and updated based on the new data the model receives [27].

Another research work on anomaly detection is proposed by [28], which is based on an HT (Hoeffding tree). It is an inductive-incremental decision-tree algorithm used for anomaly detection. A handicap of this algorithm is that it needs class labels to be available for training.

Another work to be highlighted would be that carried out by a group of Yahoo researchers [29]. Their system –called Extensible Generic Anomaly Detection System (EGADS)– allows precise, flexible, scalable, and extensible detection of anomalies, taking into account time series. The system makes it possible to separate forecasting, anomaly detection, and alerts into three separate components.

Finally, another interesting work is that contributed by [30] in which, through the integration of various technologies, the development of a disease in the leaf of a Colombian-coffee variety is evaluated and diagnosed. The project contribution relied on a model ensemble comprising four sub-models that received the data according to their nature. Once the prediction of each sub-model was made, its results were combined, calculating the weighted average. The weight of each sub-model was a value associated with its  $F_1$ -score value in the final model.

Most of the approaches to detect anomalies existing in the literature are based on models that first build a profile of what is “normal” and then point out those instances that do not fit that normal profile as anomalies (statistical methods, classification-based methods, or cluster-based methods use this approach).

A contribution of this work is to build an ensemble model that uses different algorithms that, by combining their results, will generate a new model to detect anomalies. Ensemble learning, either for classification or regression, refers to methods that generate multiple models that are combined to make a prediction [31]. Ensembles have been –extensively– used in the last decades as they are considered to provide greater accuracy and increased robustness [32]. Additionally, multiple ensemble approaches have been proposed, and several studies have reported that model diversity enhances the ensemble model’s performance as different learners generalize in different ways [33].

### III. PROPOSED METHODOLOGY

The proposed ML hybrid pipeline for real-time anomaly detection, as seen in Fig. 1, consists of two stages: i) the Manufacturing stage and ii) the Operation stage.

The *manufacturing stage* or pipeline of the Hybrid Anomaly Detection model construction process takes its name from the manufacturing process of an industrial machine. At this stage, an ML model is trained on machines’ quality control process data to validate whether the machine meets its design standards or not [34]. Thus, the objective of completing this manufacturing stage model construction task is double: (i) to use the trained model for detecting machine design/manufacturing anomalies; (ii) to later deploy it in the operation stage of the machine when it is integrated into an industrial production process, for performing a machine operation anomaly detection task. This model construction manufacturing stage is equivalent to the design phase of a classical ML workflow. The metric chosen for measuring models’ performance is the  $F_1$ -score of label  $L$ . The data set available is a slightly imbalanced (see Table 2 for class sizes percentage), where more machine’s “normal data” than “anomalous data” exists, for which the  $F_1$ -score metric is considered appropriate. The  $F_1$ -score is a value in the  $[0, 1]$  range, and it’s calculated as the harmonic mean of the estimator’s precision and recall with respect to  $L$  (see Equation (1))

$$F_1\text{-score}_L = \frac{2 \times \text{precision}_L \times \text{recall}_L}{\text{precision}_L + \text{recall}_L} \quad (1)$$

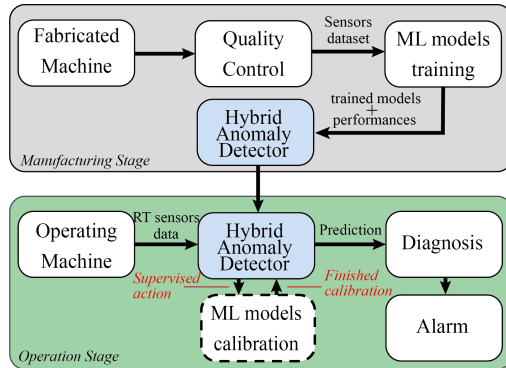
Finally, models’  $F_1$ -score ( $F_{1_i}$ ) performance ratio with respect to the sum of all  $F_1$ -scores ( $\sum_j F_{1_j}$ ) (see Equation 2) is calculated and used as the weight ( $w_i$ ) for the weighted average of the prediction done by each model multiplied by the computed weights. This weighted average assembles the Hybrid Anomaly Detection model at the manufacturing stage.

$$w_i = \frac{F_1\text{-score}_i}{\sum_j F_1\text{-score}_j} \quad (2)$$

The operation stage or pipeline refers to the phase when the machine is already running in production; in terms of a classical ML pipeline, it represents the deployment phase. Thus, this pipeline requires the machine to be able to measure the same variables taken at the manufacturing stage through industrial sensors. Once these sensors’ data are captured in real-time, they are used as inputs for the Hybrid Anomaly Detector, already trained during the manufacturing stage. This detector will diagnose based on the data received to generate an alarm for the operator in case of an anomaly. This detector can also be tuned in operation through a supervised action of the operator. If this action is triggered, the data are captured during a time window and labeled as “normal” data. The models are retrained within the hybrid anomaly detector when the data capture is complete. Once the calibration is finished, the system will be able to continue detecting anomalies in real-time.

#### A. MANUFACTURING-STAGE PIPELINE

As previously mentioned, this stage is executed when the machine is in the factory. The proposed pipeline requires that the manufactured machine goes through a quality control process [34], where sensors can capture information about the



**FIGURE 1.** Higher-level representation of the proposed Hybrid-ML pipeline for Anomaly Detection in real-time.

manufactured machine's operation during a period of time. The data captured by the sensors during the quality control process will be called *sensor data set*.

Once sensors' data are stored, the data are pre-processed for data cleaning purposes, i.e., those features that the system cannot capture with sensors when the machine is in operation are removed.

The pre-processed data are then normalized so that all features are on the same scale and comparable in later stages of the pipeline. A feature selection is then carried out to extract those variables relevant to the study; this step includes as a first filter the expert in the domain knowledge, which can give an initial selection of what variables should be maintained or discarded. Then an automatic algorithm [35] to remove redundant features is applied. Following the above, a dimensionality reduction is performed using a Principal Components Analysis (PCA) to extract the data's most representative characteristics.

The next stage is to apply a clustering algorithm, the K-means algorithm, with  $k = 2$ , which allows a distinction between a group of data samples belonging to the transient state and another group of data belonging to the steady state. To correctly label the result of the groups generated by the clustering algorithm, the cluster assigned value is first identified to the sample with the lowest timestamp of the data set. This value will correspond to the *Transient Data Group* and, therefore, all the samples containing this same cluster value will correspond to this same state. The rest of the values will be labeled as *Steady-State Data Group*.

It is also proposed for the steady-state data group to apply an outlier detection algorithm. In this case, it is proposed to use a density-based algorithm called DBSCAN, which is useful to detect outliers in applications with noise, commonly found in industrial sensor data [36].

Once the data group belonging to the transient state, stable state, and outliers (in the stable state) have been identified, a data set with new labels is generated. Furthermore, a depuration stage is carried out to obtain the final label for the data set. The transient state and outliers are labeled with a value of -1, and the normal stable data is labeled with a value of 1.

The previous data set is then divided at random and stratified into three sets: training, validation, and test. The training set corresponds to 60% of all the data, where only the normal data are used to build each ML model with cross-validation, which allows for testing its intermediate performance and tuning model hyper-parameters.

For this pipeline, the following three ML algorithms were used, selected as a result of the authors' research work on state of the art relating one-class anomaly detection for real-time systems, as they present an optimum balance of computation cost, implementation complexity, and performance [6]–[8], [12], [19]: i) LOF, which finds anomalous data points using the local deviation of a given data point to its neighbors [20]; ii) One-Class SVM (OCSVM), which finds a frontier that encloses the vast majority of data (normal data) and new upcoming data that lay outside the frontier are considered abnormal [37], [38]; and iii) Autoencoder, which reduces the input data's dimensionality by encoding the information to a smaller space. From this compressed space, it is decoded to the same dimensions as the original input. The reconstruction error in this process determines a possible anomaly [39].

Normal data are used for the training because the proposed pipeline is designed to identify anomalies based on a single class for novelty detection, and individual ML models use unsupervised algorithms.

The validation set, which corresponds to 20% of the data set, is used to obtain the definitive performance (in this case, the  $F_1$ -score value) of each trained model. The weights for the predictions of each model are then determined as the ratio of each  $F_1$ -score value (obtained using the validation set). The weights are stored to be later used for the rounded weighted average of the Hybrid Anomaly Detector component. The test set corresponds to the final 20% of the data set and is reserved for measuring the performance of the hybrid anomaly detector. The manufacturing stage pipeline is shown in Fig. 2.

## B. OPERATION-STAGE PIPELINE

This stage is executed when the machine is in operation. The operating machine generates real-time data from previously installed sensors during this process, corresponding to the same sensors used in the manufacturing stage. Each execution cycle is pre-processed and delivered to the previously obtained hybrid model, giving a diagnosis if the machine is in normal condition or if any anomalies should be reported through an alarm.

The operation stage also allows for calibrating the Hybrid Anomaly Detection models required in industrial systems that degrade over time and can be planned (e.g., every time maintenance is carried out). The operator must verify that the machine is in a stable state and under optimal conditions of normality and activate the ML models' calibration routine to carry out this process. Once this process is activated, the system will collect data during a period of time, which will depend on each system's dynamics. Each data will be stored with the normality label in the data set. This data set with normal data is then used to retrain each ML algorithm

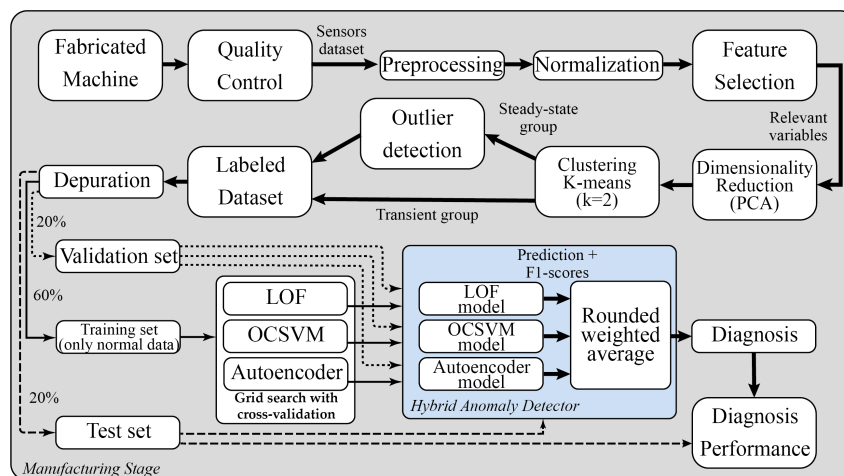


FIGURE 2. ML manufacturing stage pipeline.

TABLE 2. Air-Blowing machines’ data set characteristics.

Model version		
A	Date control	1 June 2020
	Start-End time	08:30 - 09:36
	Total Samples	1990
	Normal Samples	67.789%
	Anomaly Samples	32.211%
	Sample period	2 sec.
B	Date control	15 June 2020
	Start-End time	08:13 - 09:20
	Total Samples	2009
	Normal Samples	55.351%
	Anomaly Samples	44.649%
	Sample period	2 sec.
C	Date control	14 July 2020
	Start-End time	09:40 - 10:51
	Total Samples	2132
	Normal Samples	70.779%
	Anomaly Samples	29.221%
	Sample period	2 sec.

with cross-validation. Finally, the newly trained models are updated in the Hybrid Anomaly Detector. It should be noted that only the weights (obtained through the  $F_1$ -scores) that were acquired in the manufacturing process are used because, in the operation process, usually, there are no anomalous data to measure this performance. The operation stage pipeline can be seen in Fig. 3.

C. EXPERIMENTAL SETUP

The proposed ML Hybrid real-time anomaly detection pipeline was tested for three different industrial air-blowing machines from the local industry, with a data set generated by the quality-control process, and these machines are currently operational.

The period for collecting machines’ data is between 7 January 2020 and 2 October 2020. The data are recorded and stored at 2-second intervals. The final data set comprises 16 columns (15 variables and timestamps) with 1990 observations for Machine A, 2009 observations for Machine B, and 2132 observations for Machine C. The above-mentioned data set characteristics are shown in the table 2.

TABLE 3. Variables pre-processing at manufacturing stage.

Variable	Available at Manufacturing	Available at Operation
Flow Rate	✓	×
Nozzle Temperature	✓	×
Suction Pressure	✓	✓
Discharge Pressure	✓	✓
Flow Temperature	✓	✓
Machine Vibrations	✓	✓
RPM	✓	✓
Active Power	✓	✓
Cos Phi	✓	✓
Motor Current	✓	✓
Motor Voltage	✓	✓
Ambient Humidity	✓	✓
Ambient Temperature	✓	✓
Atmospheric Pressure	✓	×
Water Temperature	✓	×

The sensors’ data set was composed of the variables measured by sensors installed in each machine in the Quality-Control stage. The measured variables were Flow Rate, Power, Water Temperature, Nozzle Temperature, Input Pressure, Output Pressure, Flow Temperature, Machine Vibrations, RPM, Active Power, Cos Phi, Motor Current, Motor Voltage, Ambient Humidity, Ambient Temperature, Atmospheric Pressure.

The pre-processing step selects the shared variables for the manufacturing and operation stages. The variables’ pre-processing can be seen in Table 3, with a total of 11 variables selected (those with ticks in both manufacturing and operation). Additionally, samples with invalid or missing values were checked and removed from the data set in the pre-processing stage.

Afterward, the pre-processed data set was normalized to scale variables’ values, as it is recommended for data preparation in ML since some of the variables have different ranges [40]. The normalization used for this experiment was the Min-Max scaling, which scaled the data to values between 0 and 1.

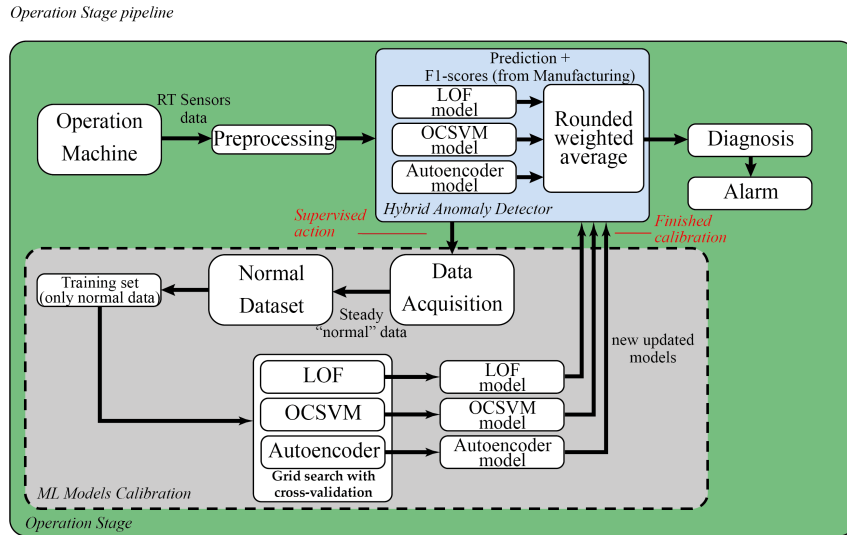


FIGURE 3. ML operation stage pipeline.

TABLE 4. Outlier detection using DBSCAN.

Parameters	Machine A	Machine B	Machine C
min_samples	7	8	8
eps	0.026	0.029	0.029
Silhouette Coef.	0.272	0.163	0.175
# of Outliers	154	200	141

The “Standard Scaler” (Z-score Normalization) was not used as the normalization method due to two main reasons: i) In the presence of outliers, the “Standard Scaler” does not guarantee balanced scales of characteristics due to the influence of outliers on the calculation of the empirical mean and standard deviation, and ii) the “Standard Scaler” assumes a normally distributed data set, which is not the case of our data set. In cases where the distribution is not Gaussian or the standard deviation is small, the “Min-Max” scaling works better [41]. Besides, “Min-Max” preserves the original distribution, does not significantly change the information embedded in the original data, and does not reduce the importance of outliers.

Following Data Normalization, a Feature-Selection step was carried out, where all the data features were validated with the expert in the domain of the machines tested. The expert determined that the “environmental” variables (Ambient Humidity, Temperature, and Atmospheric Pressure) should not be taken into account since they can present a change not necessarily related to the machine’s behavior and generate information that can disturb the final prediction of the system. The variable Cos-phi was removed because it had zero variance. Finally, the motor voltage could be explained through the motor current, and it was removed, as it was considered redundant. Finally, seven variables remained, and none of them had zero variance, so no additional variable selection step was required.

A dimensionality reduction was performed using a two-component PCA with the selected features, which

TABLE 5. Labelled data sets final samples observations.

Labels	Machine A	Machine B	Machine C
Normal	1349	1112	1509
Anomaly	641	897	623

explained the variance by 90% for each machine. A clustering was then performed using k-Means to separate the data between the Transient State and the Steady-state with  $k = 2$  groups. Furthermore, the Silhouette coefficient was used to measure the clustering’s quality, presenting a value of 0.6547 for machine A, 0.5895 for machine B presented, and 0.6744 for machine C.

Once the Transient and Steady-state data groups were separated, outliers were detected using DBSCAN in the Steady-state part. For this algorithm, two parameters called minimum samples (min\_samples) and epsilon (eps) are required, which are assigned to a list of initial values. Then the best values are found automatically to maximize the Silhouette coefficient. The list of initial values for the three machines are displayed in equations 3 and 4.

$$\text{initial\_min\_samples} = [2, 3, 4, 5, 6, 7, 8] \quad (3)$$

$$\text{initial\_eps} = [0.010, 0.011, 0.012, \dots, 0.029, 0.030] \quad (4)$$

The selected DBSCAN parameters, their performance, and the resulting number of outliers for the three machines are shown in Table 4.

Afterward, the labeled data set was created for each machine. The previously identified Transient group and Outliers are labeled as anomalies (“-1”), and the rest of the Steady-state group is labeled as normal data (“1”). The final sample observations of the three labeled data sets are shown in Table 5.



TABLE 6. Hyper-parameters selection table.

Model	Hyper-parameters	Possible Values
LOF	number of neighbours	[5, 10, 20, 30, 40]
	contamination	[0.01, 0.015, 0.02, 0.025, 0.03]
OCSVM	kernel	[linear, poly, rbf]
	gamma	[0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.5, 0.6, 0.7, 0.8, 0.9, 1]
	nu	[0.015, 0.025, 0.035, 0.045]
Autoencoder	epochs	[5, 10, 15]
	batch size	[10, 20, 30, 40]

TABLE 7. Hyperparameters and  $F_1$ -score for each generated submodel of Machine A.

Model	Hyper-parameters Value	$F_1$ -Score (Validation Set)		
		-1	1	
LOF	number of neighbours	20	0.803	0.900
	contamination	0.010		
OCSVM	kernel	poly	0.887	0.936
	gamma	0.001		
	nu	0.015		
Autoencoder	epochs	5	0.462	0.642
	batch size	30		

TABLE 8. Hyperparameters and  $F_1$ -score for each generated submodel of Machine B.

Model	Hyper-parameters Value	$F_1$ -Score (Validation Set)		
		-1	1	
LOF	number of neighbours	20	0.838	0.924
	contamination	0.010		
OCSVM	kernel	rbf	0.861	0.922
	gamma	0.001		
	nu	0.015		
Autoencoder	epochs	15	0.367	0.645
	batch size	30		

The labeled data set was then separated into three sets: 20% Validation set, 60% Training set (with only normal data), and 20% Test set, as explained in the Manufacturing stage pipeline section. For the Training set, a grid search with cross-validation was performed with five folds ( $k = 5$ ), where a set of hyper-parameters for each model was defined so that the search algorithm finds the best ones according to their respective  $F_1$ -score. These initial hyper-parameters are displayed on Table 6.

Tables 7, 8, and 9 show the selected hyper-parameters and the obtained  $F_1$ -score values for the three machines.

The last step of the proposed ML pipeline consisted of implementing an ensemble of three models: LOF, OCSVM, and Autoencoder, through a weighted average distribution. Autoencoder’s architecture is detailed in Table 10. Table 11 shows the weights for the predictions of each model, which were determined as the ratio of each  $F_1$ -score value in Tables 7, 8, and 9 with respect to the sum of all  $F_1$ -score values for each class (“-1” and “1”). As an illustrative example, for a given sample, the LOF model predicted an anomaly (-1), the OCSVM predicted normality (1), and the Autoencoder predicted an anomaly (-1) again, each output

TABLE 9. Hyperparameters and  $F_1$ -score for each generated submodel of Machine C.

Model	Hyper-parameters Value	$F_1$ -Score (Validation Set)		
		-1	1	
LOF	number of neighbours	10	0.886	0.889
	contamination	0.010		
OCSVM	kernel	rbf	0.864	0.874
	gamma	0.005		
	nu	0.015		
Autoencoder	epochs	5	0.764	0.714
	batch size	30		

TABLE 10. Autoencoder’s architecture.

N	Layer	Output Shape	# of Parameters
1	Input Layer	(None, 7)	-
2	Dense	(None, 4)	32
3	Dropout	(None, 4)	0
4	Dense	(None, 2)	10
5	Dense	(None, 4)	12
6	Dense	(None, 7)	35

TABLE 11. Weights for the predictions of each submodel.

Model	Machine	Weights (-1)	Weights (1)
LOF	A	0.373	0.363
	B	0.412	0.378
	C	0.215	0.259
OCSVM	A	0.406	0.371
	B	0.417	0.370
	C	0.177	0.259
Autoencoder	A	0.352	0.359
	B	0.344	0.353
	C	0.304	0.288

is multiplied by its respective weight, this computing the final classification of the hybrid model. Thus, considering the weights from Table 10, the output of the hybrid model will be 0.8. If this value is greater than 0, the hybrid model will classify it as a normal data point (“1”).

#### IV. RESULTS

In addition to the pipeline proposed for real-time anomaly detection, the proposed hybrid model must present improved performance metrics for the individual models. In this case, the precision, recall, and  $F_1$ -score values, as well as the Area Under the ROC Curve (AUC) of all models, were compared.

##### A. MANUFACTURING-PIPELINE RESULTS

Three machines were selected corresponding to three different model versions to check that the hybrid models worked equally well on heterogeneous equipment.

The confusion matrix allows checking which types of hits and errors (type I or false-negative errors and type II or false-positive errors) the current models have through their different metrics, such as accuracy, precision, sensitivity, and specificity. Finally, the confusion matrix of the ensemble model was analyzed to check whether it improves the individual models’ performance or not. In this respect, we focus on two metrics: i) Precision: Anomaly data are classified as normal. Also known as the False Positive Rate (FP) or Type I error. ii) Recall: Normal data are classified as an anomaly, also known as False Negative Rate (FN) or Type II error.

**TABLE 12. Machine A - confusion matrix (test set).**

Model LOF		Predicted	
Actual	Anomaly (-1)	120	39
	Normal (1)	2	237
		Anomaly (-1)	Normal (1)
Model OCSVM		Predicted	
Actual	Anomaly (-1)	130	29
	Normal (1)	4	235
		Anomaly (-1)	Normal (1)
Model Autoencoder		Predicted	
Actual	Anomaly (-1)	63	96
	Normal (1)	97	142
		Anomaly (-1)	Normal (1)
Model Hybrid		Predicted	
Actual	Anomaly (-1)	132	27
	Normal (1)	1	238
		Anomaly (-1)	Normal (1)

**TABLE 13. Machine B - confusion matrix (test set).**

Model LOF		Predicted	
Actual	Anomaly (-1)	122	31
	Normal (1)	3	271
		Anomaly (-1)	Normal (1)
Model OCSVM		Predicted	
Actual	Anomaly (-1)	137	16
	Normal (1)	23	251
		Anomaly (-1)	Normal (1)
Model Autoencoder		Predicted	
Actual	Anomaly (-1)	56	97
	Normal (1)	98	176
		Anomaly (-1)	Normal (1)
Model Hybrid		Predicted	
Actual	Anomaly (-1)	126	27
	Normal (1)	4	270
		Anomaly (-1)	Normal (1)

**TABLE 14. Machine C - confusion matrix (test set).**

Model LOF		Predicted	
Actual	Anomaly (-1)	174	46
	Normal (1)	2	180
		Anomaly (-1)	Normal (1)
Model OCSVM		Predicted	
Actual	Anomaly (-1)	163	57
	Normal (1)	5	177
		Anomaly (-1)	Normal (1)
Model Autoencoder		Predicted	
Actual	Anomaly (-1)	170	50
	Normal (1)	51	131
		Anomaly (-1)	Normal (1)
Model Hybrid		Predicted	
Actual	Anomaly (-1)	177	43
	Normal (1)	2	180
		Anomaly (-1)	Normal (1)

The Confusion matrix for machine A, machine B, and machine C are shown in Tables 12, 13, and 14 respectively.

The confusion matrix shows a generalized improvement of the hybrid model's performance compared to the other models in all three machines, both for recall and precision. For the experiments being analyzed, precision should be maximized as much as possible since it is indicative of the anomalous values detected by the system.

**TABLE 15. Machine A - metrics table (test set).**

Model	Label	Precision	Recall	$F_1$ -score	AUC
LOF	-1	0.980	0.750	0.854	0.873
	1	0.860	0.990	0.920	
OCSVM	-1	0.970	0.820	0.887	0.900
	1	0.890	0.980	0.934	
Autoencoder	-1	0.390	0.400	0.394	0.495
	1	0.600	0.590	0.595	
Hybrid	-1	0.990	0.830	0.904	0.913
	1	0.900	1.000	0.944	

**TABLE 16. Machine B - metrics table (test set).**

Model	Label	Precision	Recall	$F_1$ -score	AUC
LOF	-1	0.980	0.800	0.877	0.893
	1	0.900	0.990	0.941	
OCSVM	-1	0.860	0.900	0.875	0.905
	1	0.940	0.920	0.928	
Autoencoder	-1	0.360	0.370	0.365	0.504
	1	0.640	0.640	0.643	
Hybrid	-1	0.970	0.820	0.890	0.905
	1	0.910	0.990	0.946	

**TABLE 17. Machine C - metrics table (test set).**

Model	Label	Precision	Recall	$F_1$ -score	AUC
LOF	-1	0.990	0.790	0.878	0.890
	1	0.800	0.990	0.882	
OCSVM	-1	0.970	0.740	0.840	0.856
	1	0.760	0.970	0.851	
Autoencoder	-1	0.770	0.770	0.771	0.746
	1	0.720	0.720	0.722	
Hybrid	-1	0.990	0.800	0.887	0.897
	1	0.810	0.990	0.889	

Tables 15, 16, and 17 show the models' summary results, both individually and jointly, using their metrics for comparison.

As seen in the above tables, the performance obtained by the hybrid model improves the performance of the individual models. Thus, this justifies integrating models through a hybrid model using a weighted average improves the whole pipeline's final performance. It should also be noted that the results presented by the Autoencoder are relatively low compared to the other model; this is because the Autoencoder operates better for anomaly detection using time windows and a convolutional network architecture, which is not the case. The problem of using a convolutional architecture is that it requires time windows that could add significant delay in the operation stage and would make it difficult to compare its metrics to those of the rest of the models due to the transformation of the training, validation, and testing data that is needed to be done for being able to use the data with this type of model.

## B. OPERATION PIPELINE RESULTS

The above anomaly detection algorithm would not be useful if it could not process the trained models smoothly in a standard, real-time operation environment.

In order to measure performance, a data batch comprising 2012 samples was run for all individual models in a common computer (8GB RAM and a minimum of Intel Core i5 or equivalent; no graphic card required); the computation time needed to get the results was measured. After that, we ran the

TABLE 18. Performance results of each model in microseconds.

	LOF	OCSVM	Autoencoder	Hybrid
mean	803.6	175.4	34445.7	35982.6
std	2515.4	21.3	9999.4	11254.8
min	674.6	159.8	30300.3	31399.4
max	112896.7	446.4	174986.8	187873

same data for the hybrid model and analyzed the computation time needed to process the data. The results are presented in table 18.

As expected, the hybrid model was slower than the individual ones. Nevertheless, its time response is still over the real-time response threshold defined for a run-of-the-mill computer of 2020 (under 200 milliseconds in the worst loop of the batch analysis), thus achieving the objective established for the operation stage: real-time anomaly detection.

## V. CONCLUSION

This research work has developed and presented a Hybrid Machine-Learning Ensemble for Anomaly Detection for a Real-Time Industry 4.0 System. This ensemble consists of implementing two stages inspired by a standard industrial system: i) A Manufacturing Stage and ii) An Operation Stage. Up to our knowledge, there are no other ML methods that consider these industrial stages. The ensemble system was tested on three machines, presenting an increased  $F_1$ -score value and AUC concerning individual ML sub-models (LOF, OCSVM, and Autoencoder). The ensemble model for Machine A presented a  $F_1$ -score value of 0.904 for anomalies (-1), a  $F_1$ -score value of 0.944 for normal data (1), and an AUC value of 0.913; the ensemble model for Machine B presented a  $F_1$ -score value of 0.890 for anomalies (-1), a  $F_1$ -score value of 0.946 for normal data (1), and an AUC value of 0.905; finally, the ensemble model for Machine C presented a  $F_1$ -score value of 0.887 for anomalies (-1), a  $F_1$ -score value of 0.889 for normal data (1), and an AUC value of 0.897.

The proposed system allows vertical scaling in the number of algorithms used for the ensemble. As seen in section Results, subsection B, the hybrid model presented a maximum computation time of approximately 190 milliseconds, fast enough for real-time anomaly detection. Concerning individual models' performance, the Autoencoder results showed a low  $F_1$ -score value, so it is proposed to test other algorithms (e.g., Isolation Forest, Elliptic Envelope) to improve the overall performance of the whole assembly. However, a study of the computational cost linked to the retraining of more types of algorithms must be carried out.

Future work is proposed to study system retraining in the Operation Stage pipeline and its computational cost. It is also proposed to study the proposed system developed on machines with different levels of degradation. Additionally, a data imputation study should be carried out to generate synthetic samples for systems where some information is missing (a loss of data due to communication breakdowns is a common problem in industrial systems). Deep Learning techniques could be considered when creating

meta-classifiers using different base classifiers such as recurrent neural networks, like LSTMs, where time series need to be considered. Furthermore, a study with a larger number of machines must be carried out to see how well the hybrid model generalizes against the individual sub-models. In cases where the hybrid model does not provide any improvement, other ensemble strategies such as taking the best of the individual sub-models are considered.

Finally, as this project focuses on single-type anomaly detection, a challenge to be addressed in future work will be to be able to classify or categorize different types of faults. For that, the authors might use appropriate methods such as explainable ML or correspondingly labeled datasets.

## ACKNOWLEDGMENT

The authors would like to thank the Vicomtech Foundation for providing the necessary resources for the proper execution of this research project and University EAFIT for the research grant awarded to the principal author.

## REFERENCES

- [1] M. Xu, J. M. David, and S. H. Kim, "The fourth industrial revolution: Opportunities and challenges," *Int. J. Financial Res.*, vol. 9, no. 2, pp. 92–95, 2018.
- [2] M. Reis and G. Gins, "Industrial process monitoring in the big data/industry 4.0 era: From detection, to diagnosis, to prognosis," *Processes*, vol. 5, p. 35, Jun. 2017. [Online]. Available: <http://www.mdpi.com/2227-9717/5/3/35>
- [3] S. H. An, G. Heo, and S. H. Chang, "Detection of process anomalies using an improved statistical learning framework," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1356–1363, Mar. 2011.
- [4] A. Boukerche, L. Zheng, and O. Alfandi, "Outlier detection: Methods, models, and classification," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–37, May 2021.
- [5] J. A. Silva, E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. D. Carvalho, and J. Gama, "Data stream clustering: A survey," *ACM Comput. Surv.*, vol. 46, no. 1, pp. 1–31, 2013.
- [6] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, Nov. 2017.
- [7] V. Chandola, V. Mithal, and V. Kumar, "Comparative evaluation of anomaly detection techniques for sequence data," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 743–748.
- [8] J. Rabatel, S. Bringay, and P. Poncet, "Anomaly detection in monitoring sensor data for preventive maintenance," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7003–7015, Jun. 2011.
- [9] V. Vercauteren, W. Meert, G. Verbruggen, K. Maes, R. Baumer, and J. Davis, "Semi-supervised anomaly detection with an application to water analytics," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 527–536.
- [10] M. Fahim and A. Sillitti, "Anomaly detection, analysis and prediction techniques in IoT environment: A systematic literature review," *IEEE Access*, vol. 7, pp. 81664–81681, 2019. [Online]. Available: <https://ieeexplore.ieee.org/document/8733806>, doi: 10.1109/ACCESS.2019.2921912.
- [11] B. R. Priyanga and D. Kumari, "A survey on anomaly detection using unsupervised learning techniques," *Int. J. Creative Res. Thoughts (IJCRT)*, vol. 6, no. 2, pp. 2320–2882, 2018. [Online]. Available: <http://www.ijcrt.org/papers/IJCRT1812118.pdf>
- [12] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection for discrete sequences: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 823–839, 2012. [Online]. Available: <https://ieeexplore.ieee.org/document/5645624>, doi: 10.1109/TKDE.2010.235.
- [13] A. I. Rana, G. Estrada, M. Sole, and V. Munte, "Anomaly detection guidelines for data streams in big data," in *Proc. 3rd Int. Conf. Soft Comput. Mach. Intell. (ISCMI)*, Nov. 2016, pp. 94–98.
- [14] M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions," *Comput. Statist. Data Anal.*, vol. 52, no. 12, pp. 5186–5201, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167947307004434>

- [15] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Proc. Comput. Sci.*, vol. 60, pp. 708–713, Jan. 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050915023479>
- [16] S. Ferreira, B. Sierra, I. Irigoien, and E. Gorritategi, "A Bayesian network for burr detection in the drilling process," *J. Intell. Manuf.*, vol. 23, no. 5, pp. 1463–1475, Oct. 2012, doi: [10.1007/s10845-011-0502-z](https://doi.org/10.1007/s10845-011-0502-z).
- [17] B. Sierra, E. Lazkano, E. Jauregi, and I. Irigoien, "Histogram distance-based Bayesian network structure learning: A supervised classification specific approach," *Decis. Support Syst.*, vol. 48, no. 1, pp. 180–190, Dec. 2009.
- [18] Y. Yuan, S. Li, X. Zhang, and J. Sun, "A comparative analysis of SVM, naive Bayes and GBDT for data faults detection in WSNs," in *Proc. IEEE Int. Conf. Softw. Qual., Rel. Secur. Companion (QRS-C)*, Jul. 2018, pp. 394–399.
- [19] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, Jul. 2009, doi: [10.1145/1541880.1541882](https://doi.org/10.1145/1541880.1541882).
- [20] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, Jun. 2000.
- [21] P.-Y. Chen, S. Yang, and J. A. McCann, "Distributed real-time anomaly detection in networked industrial sensing systems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 6, pp. 3832–3842, Jun. 2015.
- [22] S. Lee, G. Kim, and S. Kim, "Self-adaptive and dynamic clustering for online anomaly detection," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14891–14898, Nov. 2011.
- [23] E. H. M. Pena, S. Barbon, J. J. P. C. Rodrigues, and M. L. Proenca, "Anomaly detection using digital signature of network segment with adaptive ARIMA model and paraconsistent logic," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2014, pp. 1–6.
- [24] S. C. Tan, K. M. Ting, and T. F. Liu, "Fast anomaly detection for streaming data," in *Proc. IJCAI Int. Joint Conf. Artif. Intell.*, 2011, pp. 1511–1516.
- [25] N. Ding, H. Ma, H. Gao, Y. Ma, and G. Tan, "Real-time anomaly detection based on long short-term memory and Gaussian mixture model," *Comput. Electr. Eng.*, vol. 79, Oct. 2019, Art. no. 106458.
- [26] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, pp. 1–39, 2012.
- [27] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, Mar. 2014, Art. no. 44. [Online]. Available: <https://dl.acm.org/doi/10.1145/2523813>, doi: [10.1145/2523813](https://doi.org/10.1145/2523813).
- [28] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2001, pp. 97–106.
- [29] N. Laptev, S. Amizadeh, and I. Flint, "Generic and scalable framework for automated time-series anomaly detection," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 1939–1947.
- [30] D. Velásquez, A. Sánchez, S. Sarmiento, M. Toro, M. Maiza, and B. Sierra, "A method for detecting coffee leaf rust through wireless sensor networks, remote sensing, and deep learning: Case study of the caturra variety in Colombia," *Appl. Sci.*, vol. 10, no. 2, p. 697, Jan. 2020.
- [31] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Comput. Surv.*, vol. 45, no. 1, p. 10, Nov. 2012.
- [32] N. Garcia-Pedrajas, C. Hervás-Martínez, and D. Ortiz-Boyer, "Cooperative coevolution of artificial neural network ensembles for pattern classification," *IEEE Trans. Evol. Comput.*, vol. 9, no. 3, pp. 271–302, Jun. 2005.
- [33] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
- [34] H. Judi, R. Jenal, and D. Genasan, *Quality Control Implementation in Manufacturing Companies: Motivating Factors and Challenges*. London, U.K.: IntechOpen, Apr. 2011, ch. 25.
- [35] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications," in *Proc. 38th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2015, pp. 1200–1205.
- [36] E. Schubert, J. Sander, M. Ester, H. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, 2017.
- [37] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.
- [38] D. M. Tax and R. P. Duin, "Uniform object generation for optimizing one-class classifiers," *J. Mach. Learn. Res.*, vol. 2, pp. 155–173, Dec. 2002.
- [39] T. Amarbayasgalan, B. Jargalsaikhan, and K. Ryu, "Unsupervised novelty detection using deep autoencoders with density based clustering," *Appl. Sci.*, vol. 8, no. 9, p. 1468, Aug. 2018.
- [40] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2018.
- [41] D. Freedman, R. Pisani, and R. Purves, *Statistics: Fourth International Student Edition* (International Student Edition). New York, NY, USA: W. W. Norton & Company, 2007. [Online]. Available: <https://books.google.es/books?id=mviJQgAACAAJ>



**DAVID VELÁSQUEZ** received the B.S. degree in mechatronics engineering from the University Escuela de Ingeniería de Antioquia (EIA), in 2011, and the master's degree in engineering from Universidad EAFIT, with emphasis on technical systems integrated design, in 2014. He is currently pursuing the Ph.D. degree in informatics with the University of the Basque Country, Spain, in collaboration with research projects from the VICOMTECH Research Center. He is also

working as an Assistant Professor with the Department of Systems and Informatics Engineering and as a Researcher with the TICs Development and Innovation Research Group (GIDITIC) and the Design Engineering Research Group (GRID), Universidad EAFIT. His research interests include adaptive systems control design, mechatronics design, industry 4.0, machine learning, computer vision, electronics optimization, embedded systems, the Internet of Things implementation, and biomedical signal processing applications.



**ENRIQUE PÉREZ** received the graduate degree in information technology engineering from the Universidad Nacional de Educación a Distancia (UNED), in 2019. He is currently pursuing the master's degree in data science with the Universitat Oberta de Catalunya (UOC), carrying out the external end of master's work in the field of artificial intelligence, developing a proposal for intelligent services for industrial blowers with the VICOMTECH Research Center, Data Intelligence

for Energy and Industrial Processes Department, through an educational cooperation agreement between the university and company. He carried out the end-of-degree project (PFG) in the field of machine learning (ML) associated with predictive maintenance in industry 4.0 environments. His research interests include the field of machine learning and deep learning (DL), creation of predictive models through advanced analytics in the Industrial Internet of Things (IIoT) systems, data visualization, and its practical application for the industry 4.0.

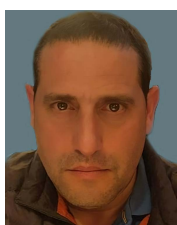


**XABIER OREGUI** received the Ph.D. degree in telecommunications engineering from the Centro de Estudios e Investigación Técnicas (CEIT), University of Navarra, more precisely in the area of electronics and communications, where he researched about multi-source virtual machine management and automatic scaling. After a small recess from research, where he spent his acquired knowledge developing educational games and "serious-games" for the company Itkasplay, in 2016, he came back to the research world in Vicomtech to the area of data intelligence and industrial processes. Back on his incorporation on Vicomtech, he has been working on multiple projects focused on data management on industrial environments using different kinds of protocols, and projects oriented on the management and analysis on big data and the visualization of that information from that same environment.



**ARKAITZ ARTETXE** received the degree in computer engineering and the M.Sc. degree in computational engineering and intelligent systems from the University of the Basque Country (UPV/EHU), San Sebastian, in 2011 and 2014, respectively, and the Ph.D. degree in computer science with emphasis on the application of knowledge engineering and machine learning to the medical domain from the University of the Basque Country, in 2017.

Since 2011, he has been working as a Researcher in the field of biomedical applications with the Technological Centre Vicomtech. Since 2018, he has been working as a Researcher with the Data Intelligence for Energy and Industrial Processes Department, Vicomtech. His research interests include machine learning, imbalanced classification, and data fusion techniques in the context of industry 4.0.



**JORGE MANTECA** is currently a Senior Industrial Engineer (specialty mechanics, machines) with ETSIIG, University of Oviedo. He is also the Technical Director of MAPNER, a company dedicated to the manufacture of machinery in the field of compressors and vacuum pumps. As main functions, he is an in-charge of technical support in commercial tasks, management and implementation of resources for the organization of the different departments of the company, and management

and coordination of the technical office, research, optimization of equipment performance, and development of new products. His previous experience includes his participation in a research project on the behavior of cryogenic fluids at CERN (European Center for Particle Physics Research), Geneva, Switzerland. He has published several scientific papers, including the publication in the international conference on cryogenics in Anchorage (Alaska), in 2003: "CONCLUSION OF THE HE SPILL SIMULATIONS IN THE LHC TUNNEL."



**JORDI ESCAYOLA MANSILLA** received the bachelor's and master's degrees in statistics and operations research, the master's (Executive) degree in business administration, and the Ph.D. degree in economics and business. He has been working in data science and predictive analytics for the last 11 years in different industries (insurance, banking, pharma, and public sector) and consulting, which includes a relevant multinational experience helping organizations and gov-

ernments. Since 2017, he has been a fellow of the prestigious organization Beta Gamma Sigma, which represents one of the highest honor institutions in business worldwide. He is currently the Consultancy Manager and the Practice Leader in data science, an Associate Professor in probability and statistics with the Universitat Politècnica de Catalunya (UPC), and an Associate Professor with the Universitat Oberta de Catalunya (UOC).



**MAURICIO TORO** received the B.S. degree in computer science and engineering from Pontificia Universidad Javeriana, Colombia, in 2009, and the Ph.D. degree in computer science from the Université de Bordeaux, France, with emphasis on artificial intelligence, in 2012. He has been a Postdoctoral Fellow with the Computer-Science Department, University of Cyprus, since 2013. Since 2014, he has been working as an Assistant Professor with the Department of Systems and Informatics

Engineering and as a Researcher with the TICs Development and Innovation Research Group (GIDITIC), Universidad EAFIT. His research interests include artificial intelligence, industry 4.0, machine learning, computer vision, and agricultural applications.



**MIKEL MAIZA** received the degree in automatic engineering and industrial electronics from the University of Mondragon, in 2000, and the Ph.D. degree from the University of York, U.K., in 2003, with emphasis on parallel computing for real-time systems. He was an External Professor with the University of Mondragon, from 2002 to 2004, an Associate Professor with the School of Engineering, University of Navarra, from 2009 to 2013, and an Associate Professor with the Department of

Applied Mathematics, University of the Basque Country, from 2015 to 2017. He has been collaborating as an External Professor with the Ecole Supérieure des Technologies Industrielles Avancées (ESTIA), since 2017. Since 2016, he has been working as a Senior Researcher with the Technological Centre Vicomtech, Data Intelligence for Energy and Industrial Processes Department. His research interests include parallel processing systems, heuristic algorithms and stochastic techniques of mathematical optimization and their integration with artificial intelligence techniques, for building stochastic models aimed at the simulation and optimization of processes and systems, and applications, such as data mining, pattern recognition, and automatic learning or early fault detection.



**BASILIO SIERRA** is currently a Full Professor with the Computer Sciences and Artificial Intelligence Department, University of the Basque Country (UPV/EHU). He is also the Co-Director of the Robotics and Autonomous Systems Group, RSAIT. He is also a Researcher in the fields of robotics and machine learning, where he is working on the use of different paradigms to improve robot's behaviors. He works as well in multidisciplinary applications of machine learning

paradigms, in agriculture, natural language processing, and medicine. He has published more than 50 journal articles, and several book chapters and conference papers.

...

# Crack Detection Method in Transport of Hygroscopic Particulate Compressed Material

David Velásquez<sup>1,2\*</sup>, Santiago Pérez<sup>1</sup>, Ricardo Mejía-Gutiérrez<sup>1</sup> and Alejandro Velásquez-López<sup>1</sup>

<sup>1</sup>Design Engineering Research Group (GRID), Universidad EAFIT, Carrera 49 No. 7 Sur - 50, Medellín, Colombia

<sup>2</sup>I+D+i on Information Technologies and Communications Research Group, Universidad EAFIT, Carrera 49 No. 7 Sur - 50, Medellín, Colombia

## Email addresses:

dvelas25@eafit.edu.co (David Velásquez), sperezca@eafit.edu.co (Santiago Pérez), rmejiag@eafit.edu.co (Ricardo Mejía-Gutiérrez), avelasq9@eafit.edu.co (Alejandro Velásquez-López)

**Abstract—** The transport of goods has been widely studied due to the importance to guarantee final product quality. The case of particulate materials is even more complicated when companies decide to innovate in the product's shape, because of the trade-off between packaging and cargo space optimization. That is the case of compressed hygroscopic particulate material, which may be addressed by compacting particles in geometric forms to improve end-user experience. However, there is a problem when transported materials are compacted particles: cracks and product damage may occur during transportation if conditions of the truck such as vehicle suspension or road conditions aren't met. These kinds of problems can be simulated to influence design decisions related to vehicle and product specifications to avoid them. This document proposes a crack identification method applied to hygroscopic particulate compressed materials subject to simulated transport conditions. An experimental approach is used to simulate package and transport conditions. Spectral analysis was used to determine if a material fulfills transport requirements to go from a given location to its destination, in terms of cracking. The article describes the experiment, data acquisition (hardware and software), as well as the theoretical basis of spectral analysis used for data processing. Finally, results are presented to explain how this analysis is capable of predicting if such a material will be damaged during transportation. The experiment considers the set of frequencies that affect the product in terms of transportation methods, compacting techniques, and packaging design.

**Index Term—** Spectral Analysis, Transportation Simulation, Hygroscopic Material, Crack Detection, Vibrations Testbed.

## I. INTRODUCTION

Some significant challenges in transport are the trade-off between packaging (for product's care) and volume (use of the cargo space), to transport more quantity of goods in a safe mode. This challenge makes the efficiency in spatial distribution one of the main concerns in product transportation. This issue is particularly relevant in the transport of particulate materials, which are packed and transported in bags. However, the air in these packages may occupy potentially available space in limited cargo space. In particulate materials, a feasible solution to optimize distribution inside the cargo space may be solved by compacting particles typically in geometric forms [1].

These kinds of compacting methods have a disadvantage: the compressed product is more fragile than its particulate form. Companies usually invest in costly special packages for preserving the product from damage [2]. Nowadays, methods are needed to provide information about the whole supply chain

and the influence of transport conditions in potential damages on transported products. If a company carries compressed material by truck, it may be relevant to have information from transport conditions to have feedback that may influence the compacting process, as well as packaging design, guaranteeing the final quality of the product.

For the compression process, variables like the pressure applied to the particulate material, compression speed, particle density, moisture can change the resistance of the final product significantly during transportation, which can be modified if the transportation conditions may crack or destroy the product [3]. These compression and transportation variables can be validated through simulations methods to prevent possible damages and guarantying product quality.

The simulation of transport conditions has been widely studied because of the importance of this phase in a product life cycle. Its importance is also linked to quality, in terms of the high possibility of product damage. These topics are the reason why simulation of transport conditions, using cost-effective methods, can offer vital information to guarantee a better quality in product transport, especially in predicting product conditions when it reaches the final destination [4].

This article is structured in five sections. After this introductory section that explained the problem, state of the art is presented to explain existing approaches and the theoretical background of the proposed method. Then the proposed method for crack detection is described from the hardware/software point of view. Section 4 presents the results and its corresponding analysis after executing the described test. Finally, in section 5, a set of conclusions are presented.

## II. STATE OF THE ART

This section is divided in two subsections: (i) Existing approaches, which presents research through the state of the art focused on existing technologies for analyzing vibrations. (ii) Theoretical background, a subsection involving the mathematics required for the proposed method using frequency analysis.

### A. Existing approaches

For simulating transport conditions using vibrations, the frequency analysis can be used to obtain required information. One of this type of analysis is the spectrogram, which can be

used to analyze speech patterns (like animal sounds) [5], radar and sonar applications for tracking targets [6], medical applications such as measuring blood flow with ultrasound information [7], and detecting cracks in cantilever beams [8]. Gillich and Praisach [9] proposed a method based on natural frequency changes for detecting damages in beam structures, using an accelerometer as sensor for measuring vibrations and concluded that natural frequency changes due to damage. Sha et al. [10] presented a novel method for single and multiple damage detection in beams using relative natural frequency changes, allowing to localize and measure the damage in a cracked beam. Onchis [11] also used frequency spectrum to identify damage in cantilever beams using a proposed procedure through Gabor transform and LASSO minimization. Sinou [12] examined the possibility of detecting the presence of open cracks in rotating machinery for low or high rotor accelerations, Webb [13] measured for the first time the full-spectral response of a Fiber Bragg Grating (FBG) sensor subjected to vibration. Yan [14] used a multi-scale enveloping spectrogram through vibration signal analysis for health diagnosis of bearings, Puchalski [15] diagnosed mechanical defects using vibrations signals and Wang [16] extracted fault features with transient vibration signal analysis. Jweeg et al. [17] investigated the effect of cracks in pipes through frequency analysis from the vibrations of the pipe, finding that the frequency decreases more and more if the crack depth is increased. Aramburo-Londoño et al. [18] presented a dynamic analysis using Finite Element Method (FEM) for the evaluation of vibration effects on hygroscopic particulate materials, where the results estimate the behavior of the compressed powder for its handling and transportation to determine the ideal conditions for the product packaging. Gomes et al. [19] proposed an experimental approach to validate standard Power Spectral Densities (PSD) through acquisition of acceleration data from electrodynamic shaker and a proposed software for signal processing. Wu et al. [20] presented a fatigue crack detection and localization technique for aluminum plates through the measurement of instantaneous baseline using a set of piezoceramic transducers and a shaker testbed. Aymerich et al. [21] investigated the effect of boundary conditions on nonlinear acoustics, which can be used for impact damage detection in composite structures. In addition to health monitoring applications, Shin [22] investigated two properties of correlation coefficients between two transient vibration signals used for the location template matching (LTM) method, which can provide an estimation of the location of an impact through vibrations signal analysis.

Most of the methods shown in literature are based on solid materials (beams, shakers, among others) damage but not on particulate compacted materials. This article presents a methodological proposal of how to combine different analysis in the domain of the frequency for processing vibrations on particulate compacted materials.

### B. Theoretical Background

The spectrogram is a graphical representation of frequency

and time information from a signal [23]. One of the ways to compute the spectrogram is processing first the Short Time Fourier Transform (STFT) [24], which can be calculated using a sliding window to divide the signal into several blocks of data. The Fourier transform calculates the analog time dependent signal into the frequency domain [25] but usually this analog signals are sampled, which requires the discrete Fourier transform (DFT) in order to translate from discrete time to frequency domain [26].

The processing required to calculate a DFT takes a lot of time. The computation of the convolution and discrete Fourier transform requires  $N^2$  operations where  $N$  is the filter length or the transform size. Using the Cooley-Tukey FFT the operations are reduced to  $N \log_2 N$  decreasing the computation time [27].

The main advantages of the FFT are the speed of computation and the memory efficiency. The DFT can be an effective process in samples of any size ( $N$ ), but it requires more computation time than the FFT and consumes more memory, because the intermediary results must be stored in all the process [28].

When FFT needs to be calculated, the algorithm pads or chops the input length ( $m$ ) to achieve the desired transform length ( $n$ ). The spectrogram applies this FFT to a  $N$ -points block of data to obtain the frequency contents of each block of data, where  $N$  is the frequency bins. The STFT aligns the center of the first sliding window with the first sample of the signal  $X$  and extends the beginning of the signal by adding zeros. The sliding window moves time steps samples to the next block of data. If the window moves out of  $X$ , it pads  $X$  with zeros. After finding the STFT, the spectrogram is calculated as the magnitude square of the elements in  $\text{STFT}(X)$  [29]. In the Figure Fig. 1 this procedure is shown.

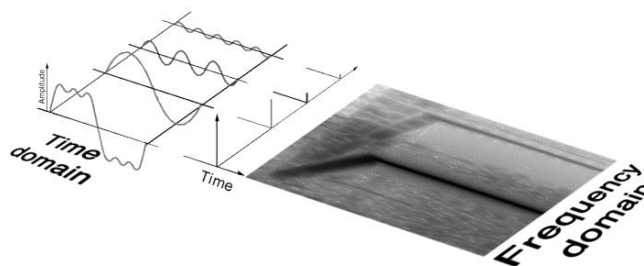


Fig. 1. STFT spectrogram calculation.

### III. PROPOSED METHOD FOR CRACK DETECTION

Our proposed method is integrated to the usual compaction and transportation process, which usually starts with the compaction of the product at a certain speed ( $P_C$  and  $S_C$  respectively). Then, the compacted products are packed in a certain envelope, which aims to protect the product from damage. The packaged products are then stacked together in the vehicle storage, exerting a pressure ( $P_p$ ) in the lower packaged products and oscillating at a frequency ( $f_T$ ). This frequency

mainly depends on the road conditions and vehicle suspension. After the vehicle arrives to the desired destination, the products are inspected to check their quality, disposing defective products. Finally, this process is repeated.

We propose a method that first verifies if a sample of the product will resist the transportation conditions, simulating them through hardware and software. The hardware part consist on a vibrations test bench, which first makes the sample oscillate at a certain frequency ( $f_T$ ) and with a simulated spring pressure ( $P_p$ ). The vibrations' data of the sample is then acquired using an accelerometer through a data acquisition device, which reports the information in a computer database. The software component consists on a developed spectrogram post-processing algorithm that checks the stored database and shows if the sample failed the test (detected crack) or if it passed the simulated transportation conditions. Figure 2 presents the contribution of this paper and how this new method connects with the usual compaction and transportation.

The following subsections will explain the hardware and software contribution details.

#### A. Hardware: Experimental test bench

A practical set of experiments was performed in order to test a proposed method based on frequency analysis for detecting cracking in fragile compressed materials during transportation. The test consisted in a package and transport simulation testbed, which applied a preload and vibrations to a compressed particulate hygroscopic material in order to see if it produces a crack in the material. Based on standard freight transport conditions [30], an oscillating frequency of 45 Hz was programmed as described in section 3.1.1 and then the resulting acceleration is acquired through data acquisition. For more details about the data acquisition refer to section 3.1.2. During this test, five different types of hygroscopic particulate

materials were used, such as powdered sugar, plaster, white cement, chocolate powdered drink and orange powdered drink to validate the proposed method. Additionally, the results will allow materials comparison in terms of behavior under the given transportation and packaging conditions. The software tool used for measuring the occurrence of a crack was a spectrogram. The expected result of this method is to detect homogeneity loss through detection of new frequency components. These components will appear when cracked pieces begin to oscillate together with the sample. This behavior implies that the sample did not resist the transportation conditions because cracks can now be detected in the spectrogram.

1) *Package and transport simulation testbed:* A variator controlled motor with an eccentric added weight was implemented to induce vibrations to the testbed. The sample is compressed with three springs that simulates the load of a pile of the same product on its top (as a pile of compressed product will do) during transport. This testbed allows the user to program an oscillating frequency to obtain acceleration orthogonal to the sample, which is processed as described in section 3.1.2. Also, the user can setup different types of geometries and sizes of compacted samples using a different set of testbed components. For the purpose of this experiment, samples with circular geometries were used but the test can be extended to other type of shapes in the samples. Figure 3 shows the diagram of the vibrations testbed with the vehicle transportation simulator.

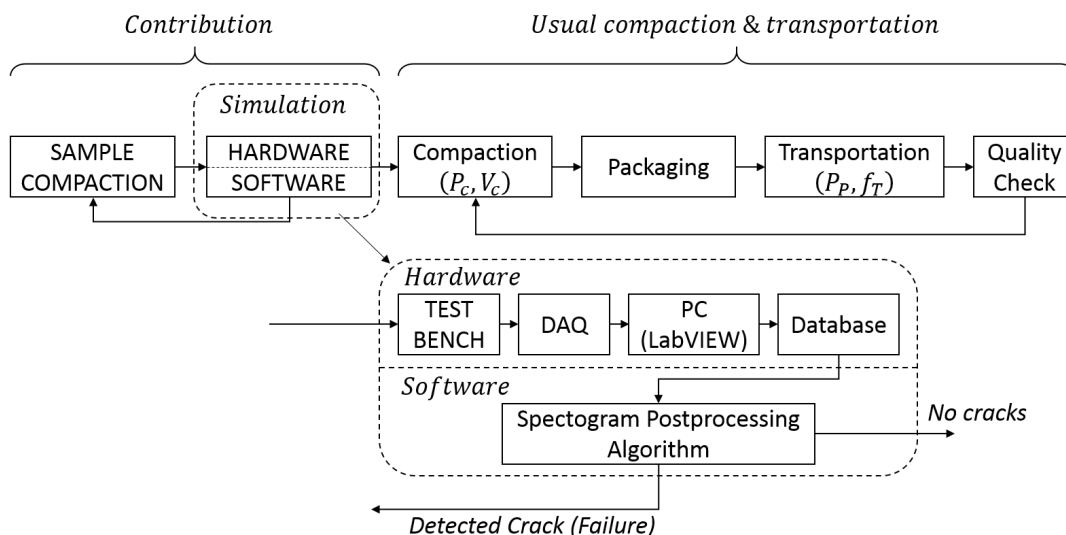


Fig. 2. Proposed new method for crack detection.



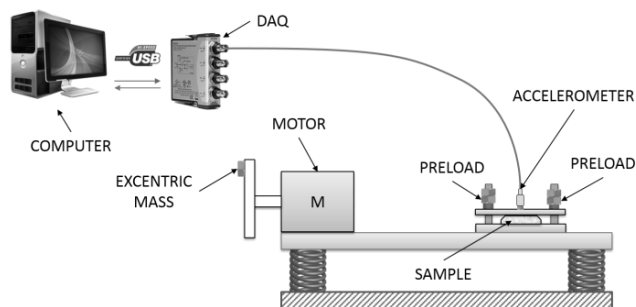


Fig. 3. Graphical explanation of package and transport simulation testbed.

2) *Data Acquisition Setup*: For acquiring the vibrations' data, a NI DAQmx 9234 was used, which is a four-channel C Series dynamic signal acquisition module for making high-accuracy audio frequency measurements from integrated electronic piezoelectric (IEPE) and non-IEPE sensors with NI CompactDAQ or CompactRIO systems. The NI 9234 delivers 102 dB of dynamic range and incorporates software-selectable AC/DC coupling and IEPE signal conditioning for accelerometers and microphones. The four input channels simultaneously digitize signals at rates up to 51.2 kHz per channel with built-in anti-aliasing filters that automatically adjust to your sampling rate.

For measuring the vibrations on the sample, an uni-axial Kistler accelerometer was implemented. This accelerometer has a measuring range of  $\pm 50$  g and a sensitivity of 99.8 mV/g.

Using the previous hardware, a communication was established with an algorithm developed in the software LabVIEW for storing the incoming data into a ".txt" file at a sampling rate of 50 kHz. Figure 4 presents the flow diagram of the algorithm, in which the acceleration is stored for each sample, then the Root Mean Square (RMS) value is calculated in order to see how much real acceleration is being applied to the vibrations testbed, then the acceleration is plotted in a graphic for visualization. When the STOP button is pressed, all data is stored in a ".txt" file for later processing.

The designed Human Machine Interface (HMI) in the software LabVIEW for data acquisition is shown in Figure 5, which contains a start and stop buttons, a sampling frequency numeric control, the time elapsed by the test, information about the acceleration of the vibrations testbed presented as plots and in a meter bar. Finally, it contains RMS acceleration value in a numeric indicator.

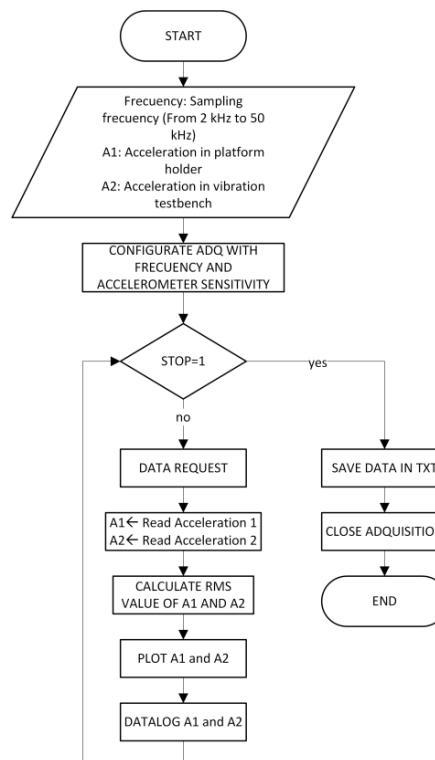


Fig. 4. Algorithm for vibrations acquisition.

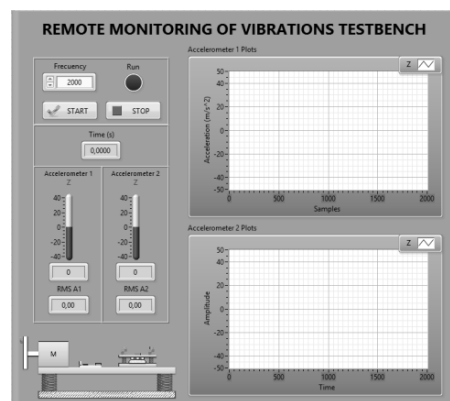


Fig. 5. LabVIEW HMI Interface for vibrations acquisition.

### B. Software: Algorithm for data processing

For analyzing the acquired vibrations' data, an algorithm was developed in MATLAB using spectrogram function plus a direct-form FIR low pass digital filter [31] for data conditioning and filtering possible electronic noise for frequencies over 100 Hz. The inputs for the low pass digital filter are shown in Table I and the output coefficients were used in the integrated function *filter* from MATLAB.

TABLE I  
LOW PASS DIGITAL FILTER INPUTS

Variable	Value
FIR Design Method	Equiripple
$F_s$	50000 Hz
$F_{pass}$	90 Hz
$F_{stop}$	100 Hz
$A_{pass}$	0.1 dB
$A_{stop}$	40 dB

The developed algorithm test processing method first initializes the sampling frequency in 50000 Hz, then the data acquisition file is imported from a “.txt” to  $Y$  variable in MATLAB. A low pass filter is then applied to the data in order to remove possible electronic noise from the signal and the resulting filtered signal is stored in the variable  $X$ . Furthermore, the Spectrogram function is applied to the filtered data  $X$  with a sampling frequency of  $F_s$ , storing the results in a vector of three positions, that contains the vectors  $S$ , which is the Fourier vector,  $F$  the different frequencies vector and  $T$  the time vector. The Fourier vector  $S$  is then converted from Cartesian representation to polar representation with the “cartesian2polar” function and results are stored in  $M$  (magnitude) and  $Th$  (angle). Finally, the magnitude  $M$ , the frequencies  $F$  and the time  $T$  are stored in a .jpg file using the “colormap” function, ending the algorithm. The Figure 6 describes the flowchart of the developed algorithm.

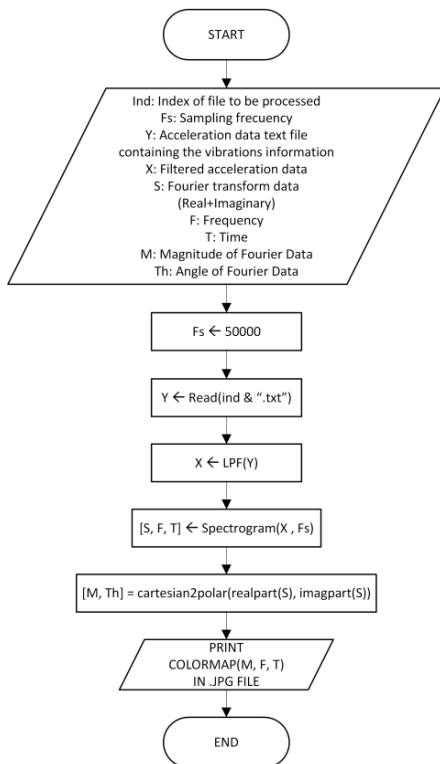


Fig. 6. Algorithm for data processing with frequency analysis.

The outputs of the algorithm are image files containing the graphical representation in a color map of the spectrogram, in

which the X-axis is the time in seconds, Y-axis is the frequency and the intensity of black is the amplitude of the vibrations signal. An example of the resulting graph of a compacted sample subject to vibrations in the testbed is presented in Figure 7 where four states can be seen: i) at the beginning when there are no oscillating vibrations, ii) a state of stabilization of the oscillating frequency, iii) the first segmentation of the sample (first crack) and iv) the total crack. The first crack can be detected when new frequencies start appearing along with the oscillating frequency.

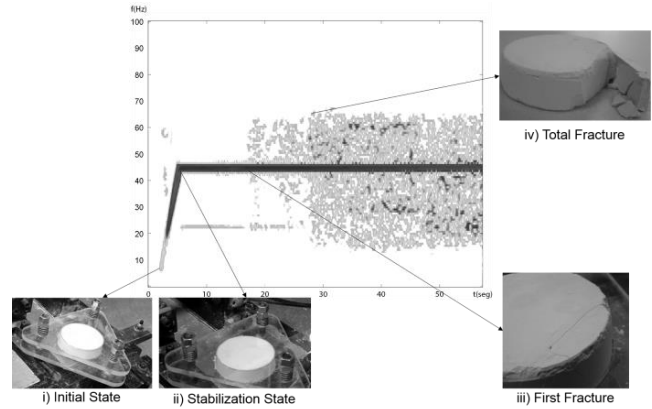


Fig. 7. Sample result spectrogram with the corresponding states.

#### IV. EXPERIMENTAL RESULTS

Four samples of different hygroscopic particulate materials were compacted with a pressure of 110 PSI. The materials of these samples were submitted to two tests of characterization:

1. Particle size test using calibrated sieve meshes (50, 100, 200 and 325).
2. Scanning electron microscope photography for the distribution of particles of the material.

Below, Table II presents the results of the particle size test.

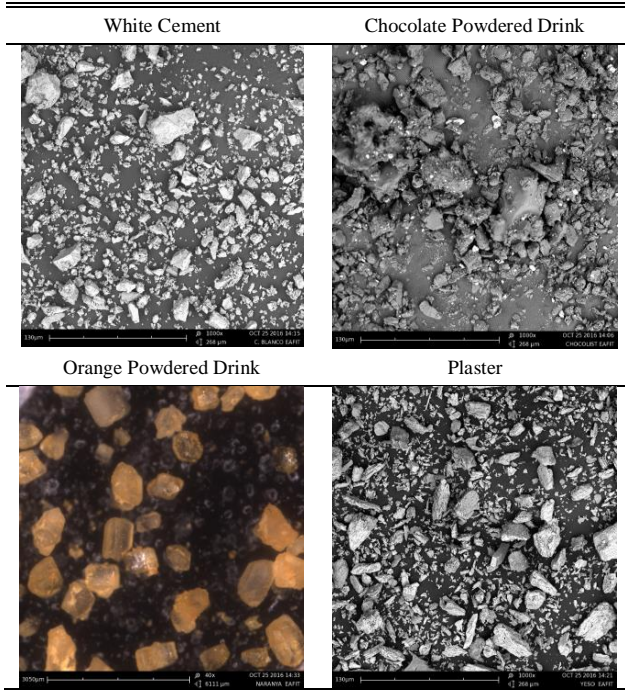
TABLE II  
PARTICLE SIZE TEST USING CALIBRATED SIEVE MESHES

Material	White Cement	Chocolate Powdered Drink	Orange Powdered Drink	Plaster
Particle Size [mm]	Percentage (%) passing the sieve			
0.300	97.83	36.65	16.37	99.70
0.150	94.16	26.90	6.80	97.41
0.075	92.99	17.79	4.22	88.86
0.045	91.24	9.86	1.39	N/A

As seen in the previous table, the White Cement and the Plaster have smaller particles than the Chocolate Powdered Drink and the Orange Powdered drink. From this variation we got different results at the compaction process, changing their mechanical properties.

At the second test, we obtained the following pictures from the scanning electron microscope at 268  $\mu\text{m}$  except the orange powdered drink, which had bigger particles that could not be zoomed at the microscope (it could scale up to 6111  $\mu\text{m}$ ).

TABLE III  
SCANNING ELECTRON MICROSCOPE PHOTOGRAPHY



From the Table III, we found that these materials have a different variation related to particle size, were the bigger particles are the support for the compaction process.

The four compacted samples were mixed according to Table IV in order to test different particles types.

TABLE IV  
INPUT DATA FOR FOUR SAMPLES OF DIFFERENT MATERIALS, ALL OF THEM AT A COMPACTION PRESSURE OF 110 PSI, SPRING LOAD OF 3 KG AND A MAIN VIBRATION FREQUENCY OF 45 HZ

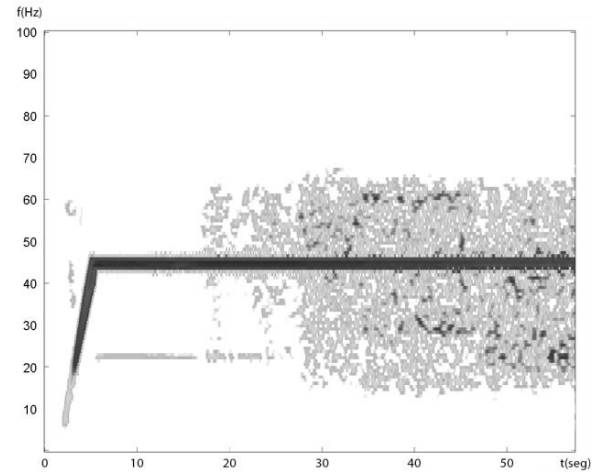
Sample #	Sample 1	Sample 2	Sample 3	Sample 4
Material	White Cement	Chocolate Powdered Drink	Plaster and Chocolate Powdered Drink	Plaster and Orange Powdered Drink

After compaction process, each sample was tested through the vibrations testbed at 45 Hz. Then the acquired data was processed using the test processing method in order to check if they will fail or not with a spring load of 3 Kg (package load simulation) and a vibration frequency of 45 Hz (truck frequency during transportation).

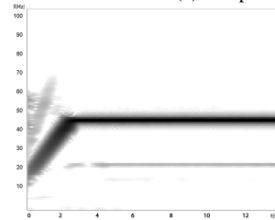
From the Figure 8, it was found that only two samples got a crack: the Sample-1 and Sample-3. This deduction was made using the method proposed in the section 3, where crack occurs when more components in frequency appear than the main frequency of vibration (45 Hz).

The Sample-1 showed the first crack at  $t = 19$  seconds and at  $t = 29$  seconds it had a total crack. This result concludes that the compressed white cement under the given conditions will have cracks during transportation. Sample-2 did not present even a minor crack, which means this compressed material can be transported without getting a crack using the previous

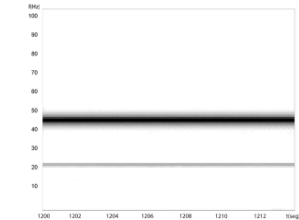
conditions. Sample-3 had the first crack approximately at  $t = 11$  seconds and after  $t = 20$  minutes, it did not present a major change which means that this type of mixture (Chocolate and Plaster) will present a minor crack during transportation under these conditions. Sample-4 showed similar results to Sample-2, it did not present even a minor crack, which means this compressed material can be transported without getting a crack using the previous conditions.



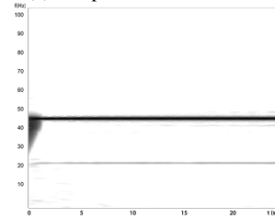
(a) Sample-1 - Crack detection.



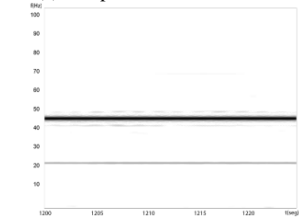
(b) Sample-2 – First 14 secs.



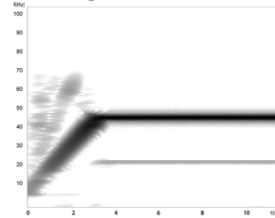
(c) Sample-2 – After 20 mins.



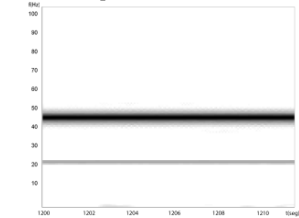
(d) Sample-3 – After 25 secs.



(e) Sample-3 – After 20 mins.



(f) Sample-4 – After 12 secs.



(g) Sample-4 – After 20 mins.

Fig. 8. 2D Grayscale Map Spectrogram.

### V. CONCLUSIONS

A method for detecting failures in products that are being transported was implemented, using the frequency analysis to verify if a crack occurred during the transportation of a given compressed product. The method considers vehicle vibrations

and all possible interactions between the compressed material, type of packaging and the vertical load applied by the compressed product over the sample (pile of the same product).

A crack can be detected in a compressed hygroscopic particulate material by finding the time in the spectrogram when new frequency components, different from the main oscillating frequency, start to appear. These new frequency components, when they are detected, indicate that some detached particles from the main sample started to oscillate around the material.

From the experimentation presented in section 4, the most resistant materials under the test conditions were the Sample-2 (chocolate powdered drink) and Sample-4 (mixture of plaster and orange powdered drink), indicating that these types of materials can be transported safely without cracks compared to the other sample materials under standard freight transport conditions.

This test is generic and it can be applied to other types of materials in a compressed form. This analysis enables to test integrity of these materials during transportation, and it allows validating compression conditions during the design of the compacted product permitting the risk reduction of cracks in transportation, improving quality of final product after transport.

It was proved that in order to have reliable results, the amount of acquired data increases proportionally to the sampling time. This requires a high computational processing capability for analyzing the spectrogram. It is then recommended to use parallel computing to reduce simulation time.

As future work, the algorithm can be improved for detecting the exact time of the failure and also vision acquisition can be implemented to the package and transport simulation testbed for comparing the spectrogram results with the vision results making the test more robust. For the purposes of this research, circular geometries in compacted samples was considered, however tests for other type of geometries in compacted samples can be further analyzed to check if it can improve the resistance of the hygroscopic compacted material for transportation.

#### ACKNOWLEDGEMENT

Authors of this article would like to acknowledge Universidad EAFIT for funding this research project. Additionally, a special thanks to the Center of Scientific Computation (APOLO) from EAFIT, for the High-Performance Computing (HPC) services used in this project. Thanks to this, it was possible to reduce the models and simulations execution time.

#### REFERENCES

[1] I. C. Sinka, "Modelling powder compaction," *KONA Powder Part. J.*, vol. 25, pp. 4–22, 2007.

[2] S. Subramonian, P. Filiccia, and J. Alcott, "Novel Soft Touch, Low Abrasion, Fine Cell Polyolefin Foams," *J. Cell. Plast.*, vol. 43, no. 4–5, pp. 331–343, 2007.

[3] O. F. Akande, M. H. Rubinstein, P. H. Rowe, and J. L. Ford, "Effect of compression speeds on the compaction properties of a 1:1 paracetamol-microcrystalline cellulose mixture prepared by single compression and by combinations of pre-compression and main-compression," *Int. J. Pharm.*, vol. 157, pp. 127–136, 1997.

[4] J. Lepine, V. Rouillard, and M. Sek, "Review Paper on Road Vehicle Vibration Simulation for Packaging Testing Purposes," *Packag. Technol. Sci.*, vol. 28, no. 8, pp. 672–682, 2015, [Online]. Available: 10.1002/pts.2129.

[5] A. V Oppenheim and R. W. Schaffer, "Spectrogram Display of the Time-Dependent Fourier transform of Speech," *Discrete-Time Signal Processing*, pp. 832–833, 2009.

[6] B. G. Ferguson, "Time-frequency signal analysis of hydrophone data," *Ocean. Eng. IEEE J.*, vol. 21, no. 4, pp. 537–544, 1996, [Online]. Available: 10.1109/48.544063.

[7] K. W. Taylor, P. N. Burns, J. P. Woodcock, and P. T. Wells, "Blood flow in deep abdominal and pelvic vessels: ultrasonic pulsed-Doppler analysis," *Radiology*, vol. 154, pp. 487–493, 1985.

[8] F. Leonard, J. Lantaigne, S. Lalonde, and Y. Turcotte, "Free-vibration behaviour of a cracked cantilever beam and crack detection," *Mech. Syst. Signal Process.*, vol. 15, no. 3, pp. 529–548, 2001, [Online]. Available: <http://dx.doi.org/10.1006/mssp.2000.1337>.

[9] G.-R. Gillich and Z.-I. Praisach, "Modal identification and damage detection in beam-like structures using the power spectrum and time-frequency analysis," *Signal Processing*, vol. 96, Part A, pp. 29–44, 2014, [Online]. Available: <http://dx.doi.org/10.1016/j.sigpro.2013.04.027>.

[10] G. Sha, M. Radziński, M. Cao, and W. Ostachowicz, "A novel method for single and multiple damage detection in beams using relative natural frequency changes," *Mech. Syst. Signal Process.*, vol. 132, pp. 335–352, 2019, doi: 10.1016/j.ymsp.2019.06.027.

[11] D. Onchis, "Observing damaged beams through their time-frequency extended signatures," *Signal Processing*, vol. 96, Part A, pp. 16–20, 2014, [Online]. Available: <http://dx.doi.org/10.1016/j.sigpro.2013.03.039>.

[12] J.-J. Sinou, "An Experimental Investigation of Condition Monitoring for Notched Rotors Through Transient Signals and Wavelet Transform," *Exp. Mech.*, vol. 49, no. 5, pp. 683–695, 2009, [Online]. Available: 10.1007/s11340-008-9193-6.

[13] S. Webb *et al.*, "Full-Spectral Interrogation of Fiber Bragg Grating Sensors Exposed to Steady-State Vibration," *Exp. Mech.*, vol. 53, no. 4, pp. 513–530, 2013, [Online]. Available: 10.1007/s11340-012-9661-x.

[14] R. Yan and R. X. Gao, "Multi-scale enveloping spectrogram for vibration analysis in bearing defect diagnosis," *Tribol. Int.*, vol. 42, no. 2, pp. 293–302, 2009, [Online]. Available: <http://dx.doi.org/10.1016/j.triboint.2008.06.013>.

[15] A. Puchalski, "A technique for the vibration signal analysis in vehicle diagnostics," *Scopus*, vol. 56, pp. 173–180, 2015, [Online]. Available: 10.1016/j.ymsp.2014.11.007.

[16] S. Wang, G. Cai, Z. Zhu, W. Huang, and X. Zhang, "Transient signal analysis based on Levenberg-Marquardt method for fault feature extraction of rotating machines," *Scopus*, vol. 54, pp. 16–40, 2015, [Online]. Available: 10.1016/j.ymsp.2014.09.010.

[17] M. J. Jweeg, E. Q. Hussein, and K. I. Mohammed, "Effects of cracks on the frequency response of a simply supported pipe conveying fluid," *Int. J. Mech. Mechatronics Eng.*, 2017.

[18] M. Aramburo-Londoño, S. Pérez-Cardona, M. Calle-Escobar, A. Velásquez-López, and R. Mejía-Gutiérrez, "Impact analysis of compressed hygroscopic particulate material," *Int. J. Mech. Mechatronics Eng.*, 2016.

[19] H. M. Gomes, D. dos Santos Gaspareto, F. de Souza Ferreira, and C. A. K. Thomas, "A Simple Closed-Loop Active Control of Electrodynamic Shakers by Acceleration Power Spectral Density for Environmental Vibration Tests," *Exp. Mech.*, vol. 48, no. 5, pp. 683–692, Oct. 2008, [Online]. Available: 10.1007/s11340-008-9134-4.

[20] W. Wu, W. Qu, L. Xiao, and D. J. Inman, "Detection and localization of fatigue crack with nonlinear instantaneous baseline," *J. Intell. Mater. Syst. Struct.*, vol. 27, no. 12, pp. 1577–1583, 2016, doi: 10.1177/1045389X15596851.

[21] F. Aymerich, W. J. Staszewski, and T. Uhl, "Effect of boundary conditions on nonlinear acoustics used for impact damage detection in composite structures," in *Health Monitoring of Structural and Biological Systems 2010*, 2010, vol. 7650, pp. 934–943, doi: 10.1117/12.847516.

[22] K. Shin, "Correlation analysis of transient vibration signals for the location template matching method," *Int. J. Mech. Mechatronics Eng.*, 2016.

[23] A. T. Catherall and D. P. Williams, "High resolution spectrograms using a component optimized short-term fractional Fourier transform," *Signal Processing*, vol. 90, no. 5, pp. 1591–1596, 2010, [Online]. Available: <http://dx.doi.org/10.1016/j.sigpro.2009.11.004>.

[24] F. Leonard, "Phase spectrogram and frequency spectrogram as new

diagnostic tools,” *Mech. Syst. Signal Process.*, vol. 21, no. 1, pp. 125–137, 2007, [Online]. Available: <http://dx.doi.org/10.1016/j.ymsp.2005.08.011>.

- [25] A. V Oppenheim, A. S. Willsky, and S. H. Nawab, “The Continuous-Time Fourier Transform,” *Signals and Systems*. p. 284, 1996.
- [26] A. V Oppenheim, A. S. Willsky, and S. H. Nawab, “The Discrete-Time Fourier Transform,” *Signals and Systems*. pp. 358–361, 1996.
- [27] P. Duhamel and M. Vetterli, “Fast fourier transforms: A tutorial review and a state of the art,” *Signal Processing*, vol. 19, no. 4, pp. 259–299, 1990, [Online]. Available: [http://dx.doi.org/10.1016/0165-1684\(90\)90158-U](http://dx.doi.org/10.1016/0165-1684(90)90158-U).
- [28] M. Chugani, A. Samant, and M. Cerna, *LabVIEW Signal Processing*, 1st ed. New York, USA: Prentice Hall, 1998.
- [29] National Instruments, “STFT Spectrograms VI.” 2008.
- [30] J. Singh, S. P. Singh, and E. Joneson, “Measurement and analysis of US truck vibration for leaf spring and air ride suspensions, and development of tests to simulate these conditions,” *Packag. Technol. Sci.*, vol. 19, no. 6, pp. 309–323, 2006, [Online]. Available: 10.1002/pts.732.
- [31] J. O. Smith, “Introduction to Digital Filters with Audio Applications,” 2007.



**Alejandro Velásquez-López** was born in Medellin, Colombia in 1978. He received his B.S. degree in Mechanical Engineering from Universidad EAFIT (Medellin, Colombia) in 2001, and his M.Sc. in Mechatronics from University of Applied Sciences (Ravensburg-Weingarten, Germany) in 2006. He works as a researcher of the Design Engineering Research Group (GRID) at Universidad EAFIT and is the manager of the postgraduate program Technical Systems Integrated Design. His research interests include mechatronic product design and solar energy

systems.



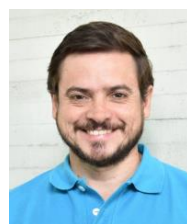
**David Velásquez** received B.S. degree in Mechatronics Engineering from university Escuela de Ingeniería de Antioquia (EIA), in 2011. He got is master degree in engineering at Universidad EAFIT with emphasis on Technical Systems Integrated Design in 2014. He works as an Assistant Professor in the Department of Systems and Informatics Engineering and as a researcher of the TICs Development and Innovation Research Group (GIDITIC) and the Design Engineering

Research Group (GRID) at Universidad EAFIT. His research interests include adaptive systems control design, mechatronics design, industry 4.0, machine learning, computer vision, electronics optimization, embedded systems, internet of things implementation and biomedical signal processing applications. Also, he is doing a Ph.D. in Informatics, at the University of the Basque Country (Spain) in collaboration with research projects from the VICOMTECH research center.



**Santiago Pérez**, B.S. degree in Mechanical Engineering from Universidad EAFIT (Medellin, Colombia), with several studies ranging from business, innovation and entrepreneurship to solid rocket motor design; Santiago Leads INMOTION GROUP [8 years in operation], a Spin-Off at EAFIT University, conformed inside the Design Engineering Research Group (GRID) specialized in sustainable transportation and mobility optimization. His research interests include data science, analytics,





computational data processing, electric vehicles and mobility optimization.



**Ricardo Mejía-Gutiérrez** is Full Professor at EAFIT University (Colombia) since 2009. He is Director of the Design Engineering Research Group (GRID) and Associate Editor of the Springer’s International Journal on Interactive Design and Manufacturing (IJIDeM). He worked from 1999 to 2008 in recognized research centers from France, Germany and Mexico. He holds a PhD in Mechanical Engineering (ECN, France, 2008), M.Sc. in Manufacturing Systems (ITESM, Mexico, 2003) and BSc in Production Engineering (EAFIT, Colombia,

2000). His current research interests include: Sustainable Design, Electric Mobility, knowledge Engineering, Integrated Product Development, Design methods & tools, Collaborative/Concurrent Engineering, Product Lifecycle Management and Manufacturing Technologies.

# EDAR 4.0: Visual-Analytics for Waste Water Management

David Velásquez , Paola Vallejo, Mauricio Toro , Juan Odriozola, Aitor Moreno, Gorka Naveran, Mikel Maiza  and Basilio Sierra 

**Abstract**—Waste-Water Treatment-Plants (WWTPs) operations manage a massive amount of data that can be gathered with new Industry 4.0 technologies such as the Internet of Things and Big Data. These data are critical to allow the wastewater-treatment industry to improve its operation, control, and maintenance. However, the data available needs to be improved and enriched, partly due to its high dimensionality, low reliability, and the lack of appropriate data analysis and processing tools for such systems. This paper presents a visual-analytics-based platform for WWTP that allows users to identify relationships among data through data inspection. The results show that the tool developed and implemented for a full-scale WWTP allows operators to construct Machine Learning (ML) models for water quality and other water-treatment process variables. Thus, plant operation scenario analysis and optimization can be performed. The validation of the variables influencing the created ML models by domain experts proved their appropriateness.

**Index Terms**—Waste water management, visual analytics, industry 4.0, data driven modeling, waste water treatment plant (WWTP)

## I. INTRODUCTION

NEWLY-connected industry objects are generating data at a fast speed that must be stored, processed, and monitored—in real-time—to make decisions that optimize the new Industry 4.0 factories' production. Numerous challenges

involve this newly-generated data and how to visualize it, such as reducing dimension and visualizing multivariate real-time data.

An advanced data processing and visualization approach that can be followed is *visual analytics*. Keim et al. [1] defined *visual analytics* as a combination of automated analysis techniques with interactive visualizations for an adequate understanding, reasoning, and decision-making based on extensive and complex data sets. Visual analytics focuses on creating new tools that enable users to: i) synthesize information that allows getting new insights from massive heterogeneous sets of data, ii) detect the current states of systems and discover possible new states, iii) provide real-time assessments and perform actions based on these assessments.

Keim et al. [1], in 2008, proposed six challenges for visual analytics: i) scalability with large data volumes and dimensionality, ii) graphical representation of data quality, iii) visual representation of levels of detail, iv) new display interfaces such as large-scale power walls, v) evaluation frameworks for visual analytics, vi) and refreshing the interactions in real-time (e.g., less than 100 ms). Many of these issues still need to be solved today.

Recently, Diez-Olivan et al. [2] found that a new challenge is using visual analytics for enhanced understandability in the context of Industry 4.0. This challenge is one handicap for the widespread adoption of data-based analysis is the industrial plant operator's assimilation of information. According to Diez-Olivan et al. [2], when it comes to data analysis, the information produced by the deployed models cannot be processed straightforwardly by non-specialized personnel unless some preprocessing is conceived for an improved, more intuitive understanding of the captured patterns.

Successful Waste Water Treatment-Plants (WWTPs) can be managed by seeking optimal process conditions and identifying essential factors, features, or patterns for data-supported decision-making. Newhart et al. [3] highlighted that WWTP operators usually store a sufficiently large amount of historical data. Also, recent advancements in data-driven process control and performance analysis and more substantial computation power “could provide the wastewater treatment industry with an opportunity to reduce costs and improve operations” [3]. However, the limited investments in instrumentation, control, and automation of WWTPs and the need for data-science background for WWTPs professionals are limitations to making the best of the data.

Manuscript received January 9, 2023; revised Month X, 2023; accepted Month Y, 2023. Date of publication Month Z, 2023; date of current version Month Z, 2023. This work was supported in part by the University EAFIT, and the Vicomtech Foundation under the project EDAR 4.0. (Corresponding author: David Velásquez).

David Velásquez, Paola Vallejo, and Mauricio Toro are with the RID on Information Technologies and Communications Research Group, Universidad EAFIT, Carrera 49 No. 7 Sur - 50, Medellín, Colombia (e-mail: dvelas25@eafit.edu.co; pvallej3@eafit.edu.co; mtorobe@eafit.edu.co).

Juan Odriozola, Mikel Maiza, and David Velásquez are with the Department of Data Intelligence for Energy and Industrial Processes, Vicomtech Foundation, Basque Research and Technology Alliance, 20014, Donostia-San Sebastián, Spain (e-mail: jodriozola@vicomtech.org; mmaiza@vicomtech.org).

Aitor Moreno is with the Department of R&D, Ibermática, Cercas Bajas, 7 int. – Office 2, 01001, Vitoria-Gasteiz, Spain (e-mail: ai.moreno@ibermatica.com).

Gorka Naveran is with the Department of R&D, Giroa-Veolia, Laida Bidea, Building 407, 48170, Zamudio, Spain (e-mail: gorka.naveran@veolia.com).

Basilio Sierra is with the Department of Computer Science and Artificial Intelligence, University of Basque Country, Manuel Lardizabal Ibilbidea 1, 20018, Donostia-San Sebastián, Spain (e-mail: b.sierra@ehu.eus).

TABLE I  
WATER QUALITY REQUIREMENTS FROM EUROPEAN DIRECTIVE  
91/271/EEC

Variable	Absolute Values	Performances
$BOD_5$	25 mgO <sub>2</sub> /L	70%
$TCOD$	125 mgO <sub>2</sub> /L	75%
$TKN$	10 mg/L	90%
$TP$	1 mg/L	80%
$TSS$	35 mg/L	70%

One of the most critical factors affecting the Big Data era's decision-making process is finding relevant data and getting meaningful information from them. To address this problem in the context of WWTPs, project *Estación Depuradora de Aguas Residuales* (EDAR 4.0) aims to develop a set of WWTP operation and management systems by combining (i) cloud computing, (ii) data intelligence, and (iii) visual analytics. EDAR 4.0 aims to provide greater data storage, processing, computation, and decision-making capabilities for WWTP operation [4]. EDAR 4.0's results were tested and validated in a full-scale municipal WWTPs: La Cartuja (Zaragoza, Spain), operated by Veolia.

Five variables related to WWTP's operation and management were analyzed in EDAR 4.0: Biological Oxygen Demand-5 ( $BOD_5$ ), Total Chemical Oxygen Demand ( $TCOD$ ), Total Kjeldahl Nitrogen ( $TKN$ ), Total Phosphorous ( $TP$ ), and Total Suspended Solids ( $TSS$ ). These variables are not selected randomly but are the variables that the European Directive 91/271/EEC establishes as quality requirements to be fulfilled at the effluent of a WWTP. Likewise, in the case that applies, as the WWTP is located in a region (Aragon, Spain) declared as an area sensitive to eutrophication, specific values for total phosphorus and total nitrogen are applied. Table I shows the quality requirements taken as a reference in this project based on the previously mentioned European Directive.

This paper presents a platform that allows the creation of data-based models for the simulation, prediction, and optimization of WWTPs.

Two modules compose this platform: i) a module for the monitoring and prediction of water quality, and; ii) a module for the creation of water quality and energy management, ML models and subsequent future scenario analysis and optimization of the WWTP.

In what follows, a brief state-of-the-art is presented in Section II. Then, the methodology is shown in Section III and the results in Section IV. Finally, conclusions and future-work directions are proposed in Section VI.

## II. STATE OF THE ART

The state-of-the-art is divided into three parts. The first part presents research on model-based wastewater management. The second part summarizes different works on visual analytics. Finally, the last part explains research on data-based wastewater management.

### A. Model-based wastewater management

A brief state-of-the-art on wastewater treatment plants modeling based on *Ordinary Differential Equations (ODEs)* is

presented in what follows.

The most common approach to optimize the process operation against fluctuating influent water quality is to apply process control and simulation of the process for deriving the optimal-operation method. For process simulation, ODEs have been widely used. To simulate WWTPs using ODEs, it is essential to first model the process's steady-state under a given set of disturbances and operating conditions. However, a disadvantage of this is that the calculation time is extended when analyzing the ODEs. Recently, Jong-Rack et al. [5] proposed an improved Newton-Raphson method to shorten the computation time. The above shows that there is still active research on the simulation of wastewater treatment plants using ODEs.

In another work, Flores-Alsina et al. developed a plant-wide aqueous-phase chemistry model describing pH variations interfaced with industry-standard models [6]. Flores-Alsina et al. formulated the general equilibria as a set of Differential-Algebraic Equations (DAEs) instead of ODEs to enhance simulation speed. Additionally, Flores-Alsina et al. applied a multi-dimensional version of the Newton-Raphson algorithm to handle multiple algebraic inter-dependencies.

It is important to mention that the International Water Association (IWA) benchmark simulation model has been available for several years to create platforms for control strategy benchmarking of activated sludge processes. In 2006, Jeppsson et al. extended the IWA benchmark to facilitate control-strategy development and performance evaluation at a plant-wide level and, consequently, includes both pre-treatment of wastewater and the processes describing sludge treatment [7].

Finally, the work by Li et al. [8] did not involve WWTPs but is worth mentioning because it presents a combination of ODEs with ML. Their paper presents a Fourier Neural Operator for modeling turbulent flows with zero-shot super-resolution. This work showed higher speed and better accuracy compared to classical solvers.

### B. Visual Analytics

Visual Analytics combines interactive visualizations with data analysis and machine learning (ML) to empower people to analyze, explore, and understand large data [9]. The Visual Analytics process can be generalized by the framework proposed by Van Wijk [10] (see Fig. 1). The first step is acquiring data stored in a database or from a data stream. This data is then analyzed and processed to extract the most critical features presented in the visualization stage. An image is then generated during the visualization stage, representing this processed and selected data or by the user's specifications. Afterwards, the user sees this image, perceives it, and the user will generate insights and knowledge from the recent image. This stage may be repeated as long as the user looks through the whole image. Finally, the user may generate hypotheses, which will be detailed through an exploration and analysis stage. Furthermore, a new analysis may be required, translating into a specification stage, where the user can interact with the current visualization to generate new knowledge.

Yuan et al. [11] presented a survey of visual analytics for ML before, during, and after model building. Before model

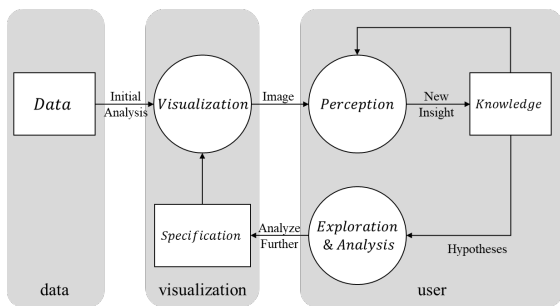


Fig. 1. Visual-Analytics process framework adapted from [10].

building, visual analytics is used to improve data quality and feature quality. Visual analytics is then used during model building to facilitate model understanding, diagnosis, and steering. After model building, visual analytics is used to understand static and dynamic data analysis results.

According to Diez-Olivan *et al.* [2], visual analytics has emerged as a promising discipline to visually adapt the discovered insights and optimally present results to different human profiles. These aspects are essential in real-use cases to deploy models for data analysis in industrial plants with minimum usability and practical utility guarantee.

As an example of visual analytics, Li and Ma [12] proposed P6, a declarative language for rapidly specifying the design of visual-analytics systems that integrate ML and visualization methods for interactive visual analysis. P6 was motivated by three goals: interactive ML and visualization (to facilitate automated analysis), interactive and scalable systems (to process and visualize large datasets), and declarative visual analytics (to create interactive visualization applications). In P6, the specification's basic unit is a pipeline composed of specifications: data, analysis, view layout, visualizations, and interactions.

Nawaz *et al.* [13] developed an intelligent Human-Machine Interface (HMI) called ANKSyst that allows operation and decision support for the ANaerobic AMMonium OXidation (ANAMMOX) process in WWTPs. This tool integrates soft sensing, decision-making, and model simulation for supervisory control, which consists of an Artificial Neural Network, a Kalman filter, and a principal component analysis algorithm.

Additionally, Li *et al.* proposed that the declarative specification for visual analytics allows non-specialists to develop advanced data analytics and communication solutions that combine the best of human and artificial intelligence [12]. According to Endert *et al.* [14], Visual analytics systems combine machine learning (or other analytic techniques) with interactive data visualization to facilitate insight and analytical reasoning. Endert described three categories of models and frameworks: (i) models meant to describe the people's cognitive stages for analyzing data; (ii) models and frameworks that describe interaction and information design of visual analytic applications; and (iii) ML frameworks that emphasize the importance of training data and ground truth to generate accurate and effective computational models. Endert and Keim *et al.* [15] mentioned that the most common ML algorithms

used with visual analytics are: (i) dimension reduction, (ii) clustering, (iii) classification, and (iv) regression.

As stated by Liu *et al.* [16], "interactive model analysis, the process of understanding, diagnosing, and refining an ML model with the help of interactive visualization, is very important for users to solve real-world artificial intelligence and data mining problems efficiently." Liu *et al.*'s paper presents a classification of relevant work in visual analytics into three categories: (i) understanding, (ii) diagnosis, and (iii) refinement. Liu highlights that many techniques generate static images to indicate which parts of an image are most important to the classification. However, interactive visualization plays a critical role in model understanding and analysis to help people gain insight into various ML models. For that reason, our proposal addresses the dynamic creation of demand-driven models, for example, a water-quality model, and how its response helps to understand a particular variable.

Massive data sets and complex, long-running analytics are common in various domains. Stolper *et al.* [17] introduced the Progressive Visual Analytics (PVA) concept. PVA is a workflow to provide the user with meaningful intermediate results if the final result's computation is too costly. Based on these intermediate results, the user can visualize, analyze, and interpret partial results before obtaining the complete results.

Visual analytics, in the industrial context, has been used widely. Jonker *et al.* [18] followed a visual analytics approach to aid this deep understanding of complex time-series models as an application to economic data. Sun *et al.* [19] proposed PlanningVis, a visual analytics system to support the exploration and comparison of production plans with three levels of details: a plan overview presenting the overall difference between plans, a product view visualizing various properties of individual products, and a production detail view displaying the product dependency and the daily production details in related factories. Finally, Wu *et al.* [20] reported the design and implementation of an interactive visual analytics system, which helps managers and operators at manufacturing sites leverage their domain knowledge and apply substantial human judgments to guide the automated analytical approaches, thus generating understandable and trustable results for real-world applications. Our system integrates advanced analytical algorithms (e.g., Gaussian mixture model with a Bayesian framework) and intuitive visualization designs to provide a comprehensive and adaptive semi-supervised solution to equipment condition monitoring.

### C. Data-based wastewater management

In WWTPs, visual analytics facilitates rapid and interactive exploration of multiple views of the same high-dimensional data. It is possible to have a global view of data behavior through different colors, orientations, and data. Interactive visualization of trade-offs in multiple dimensions is well-suited for situations where stakeholders have diverse interests [21].

Recently, Kim *et al.* proposed an operator decision support system (ODSS) to support WWTPs operators in making appropriate decisions [22]. Kim *et al.*'s system accounts for water-quality variations in the WWTP and comprises two



diagnosis modules, three prediction modules, and a scenario-based supporting module. The prediction modules are based on the k-nearest neighbors (k-NN) method to forecast water quality three days in advance.

Similarly, Heo et al. [23] proposed a hybrid influent forecasting model based on multimodal and ensemble-based deep learning. This tool predicts a WWTP's long-term (daily) and short-term (hourly) influent load.

In WWTPs, the portion of operating costs related to electric power consumption is increasing. Piao et al. used mathematical modeling to deduce six improvement plans to reduce electric power consumption [24]. The electric power consumption for Piao et al.'s suggested plans was estimated using an artificial neural network.

### III. METHODOLOGY

The methodology followed in this article is inspired by the proposal of [25], who argued that these are the typical steps in successful data analysis and mining:

- 1) Data collection and acquisition. It is the process of gathering and measuring information on targeted variables; it is divided into the following activities:
  - a) Analysis of data origin and frequency.
  - b) Quantification of data uncertainty.
  - c) Compilation of data from various sources.
- 2) Data management and data validation. It checks source data's accuracy and quality before using, importing, or otherwise processing it. It is composed of the following activities:
  - a) Definition of erroneous data.
  - b) Detection and removal of outliers based on the variable analysis.
  - c) Detection of outliers based on physical processes.
- 3) Data visualization. It is the graphical representation of information and data; its main activities are:
  - a) Exploration and visualization of data.
  - b) Development of intuitive, powerful visualizations.
  - c) Development of algorithms for the prediction of future conditions.

AvRuskin et al. states that "due to the physical nature of wastewater process data, it is recommended that laboratory, operations, and engineering staff be consulted at all points in the process to confirm assumptions" [25].

According to Anderberg, cluster analysis can be used to develop inductive generalizations [26]. Clustering analysis has been used in the domain of water quality to i) investigate the spatiotemporal structure of determinants in a set of 21 Scottish lakes [27], ii) evaluate the water quality of three different cross-sections of the Fen River [28], and iii) evaluate the quality of underground water [29].

Radar plots are a useful way to present multivariate data. According to Joan Saari [30], "radar plots have great utility in situations in which there are large numbers of independent variables, possibly with different measurement scales". In addition, Joan Saari found that "radar plots have a particular relevance for researchers who wish to illustrate the degree of

multiple-group similarity/consensus or group differences on multiple variables in a single graphical display" [30].

### IV. PROPOSED EDAR 4.0 TOOL

EDAR 4.0's architecture has the WWTP process as the base, which includes a factory-level data acquisition of all the processes that make up a WWTP. This process can be classified into three main standard sub-processes. First, the influent represents the entry of the incoming water and its preliminary and primary treatment, usually performed in a primary settling or sedimentation tank. Second, the biological treatment process is the central part of the so-called secondary treatment. It represents the biological wastewater treatment process of different types of bacteria and protozoa, which can be complemented by additional chemical treatments. Third, the effluent process represents the wastewater treatment plant output. This output receives directly treated water or water that goes through a secondary decantation or sedimentation tank, which can also be considered part of the plant's secondary treatment.

The processes and sub-processes of a WWTP are generally controlled by one or more programmable logic controllers (PLC) integrated with different sensors and actuators. All control information is displayed locally via human-machine interfaces (HMIs), usually integrated into a SCADA (Supervisory Control And Data Acquisition) system. All the information on the system is generally shared on a local network (LAN) based on an industrial protocol.

In EDAR 4.0, this is extended to a 4IR system architecture by establishing an additional cloud-based IoT infrastructure that can be reached via the internet, so the overall WWTP and its ICT infrastructure must have secure access. In this cloud, various services are integrated, such as WWTP monitoring, cloud-based IoT data acquisition and storage, information visualization, data analysis, and related services, such as visual analysis and scenario analysis for plant operation optimization through machine learning models.

An example of accessing the above IoT cloud infrastructure and related services is via the HTTP REST protocol. An example of a data analytics service is to classify different types of water quality and predict (forecast) how water quality will change over time. Finally, with the above IoT cloud platform running, the data from the sewage treatment plant can be displayed on a webpage where remote users can execute water quality analysis and other plant monitoring functionalities. Figure 2 details a view of the EDAR 4.0, 4IR system architecture. This figure also explains the software tools used for the IoT cloud components. The Python-based Flask library's API was used in this work. For data storage, a PostgreSQL database was used. The software Rapidminer was used for data analytics and ML-based model construction. Finally, for the visualization part, the Bokeh library was used. The following subsections detail each of the ML modules developed.

#### A. Water-quality monitoring

The dataset obtained from the "La Cartuja" WWTP SCADA system was subjected to a series of steps to preprocess it and

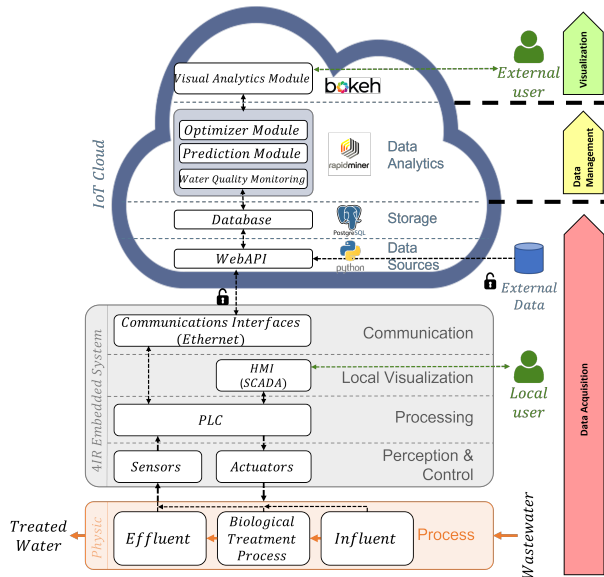


Fig. 2. EDAR Architecture.

leave it ready for the Data Cleaning process. Once the data has been cleaned, a PCA is applied to extract the two main components that define the dataset. Furthermore, a clustering process is executed using the K-means algorithm with  $k=4$ , where each group identified by the algorithm belongs to a water quality cluster.

The platform allows parameterizing if the water quality monitoring is displayed according to the water treatment’s contaminants removal performance or effluent’s water quality absolute values, in the frontend. Another parameter that the user could set from the platform is the WWTP operation period. The above was implemented because the “La Cartuja” wastewater treatment plant had a plant design and equipment improvement over time, so it was essential to monitor and separate these two periods. Water quality profiles (or clusters) are plotted using a line profile chart and a spider chart. Figure 3 displays the monitoring module of the EDAR 4.0 platform. This plot shows that the worst water quality is the blue-colored cluster (Cluster 0), whereas the best quality is the red-colored one (Cluster 3). Besides, it can be noted that the WWTP should improve the treatment of the NTK chemical variable.

**B. Water-quality prediction**

The water-quality prediction tool predicts the number of times the WWTP could have each water quality cluster in a month. For that, the backend implements a Holt-Winters time-series forecasting. Two plots are displayed in the frontend: i) The time-series cluster prediction plot and ii) The outlier probability plot. These plots can be seen in Figure 4. The vertical dashed line separates the real dataset (monitoring) from the prediction data. WWTP operators should ideally see in this graph that the highest prediction count is in the best water quality, the red cluster (Cluster 3), and the lowest count in the worst water quality, the blue cluster (Cluster 0).

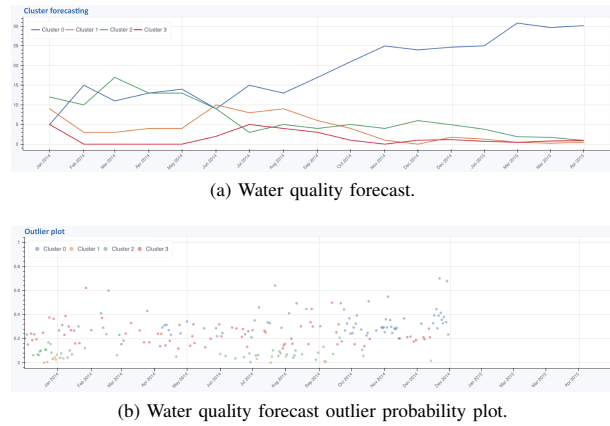


Fig. 4. Visual Analytics Water Quality Prediction Platform.

**C. WWTP Model Creation & Simulation**

At this stage, it is possible to create a data-based model for any WWTP process variable, including energy, water quality, or process operation and control-related variables of the wastewater treatment process. The model created in the platform by default is a water quality model. However, other process variables, such as energy consumption (kilowatts per day), can be modeled as a function of other process variables. In the back end, the machine learning system implemented can detect the most relevant variables for the models to be developed based on information such as a process variables’ correlation matrix. The method selected for the creation of models is based on decision trees. Once the model is created, it is possible to interact with the platform’s variables relevant to that model. Once the values are selected, a prediction of modeled variables’ range of values can be performed with those values with which the model is simulated. This process is shown in Figure 5, which is based on an example of modeling electricity consumption; a set of values is given for the relevant variables, and after running the simulation, the platform predicts that the WWTP will be at a range1 ( $-\infty$  to 59816 kW) of energy consumption.

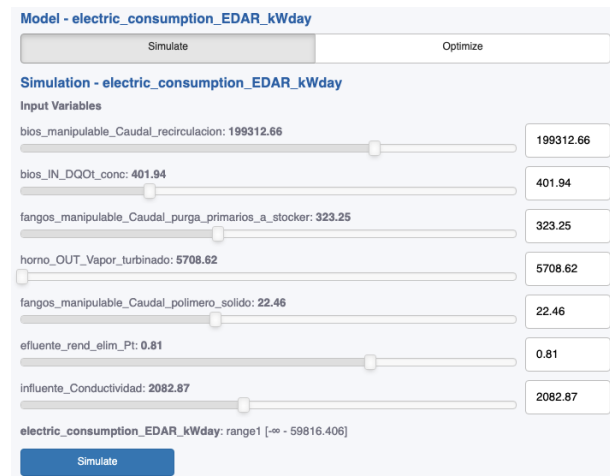
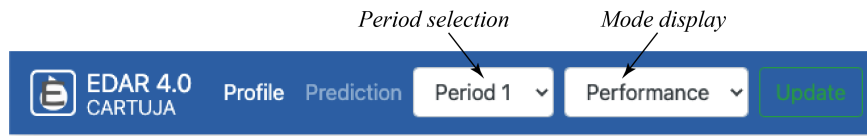
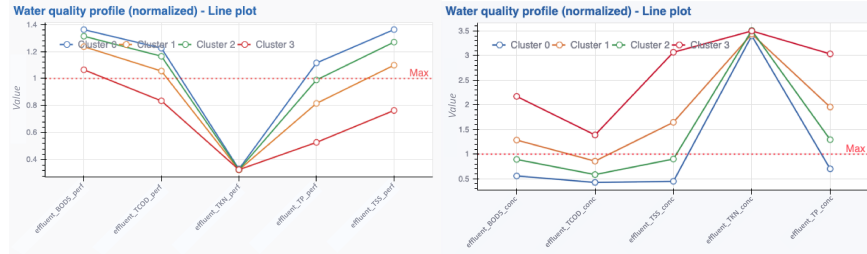


Fig. 5. Energy consumption model simulation.

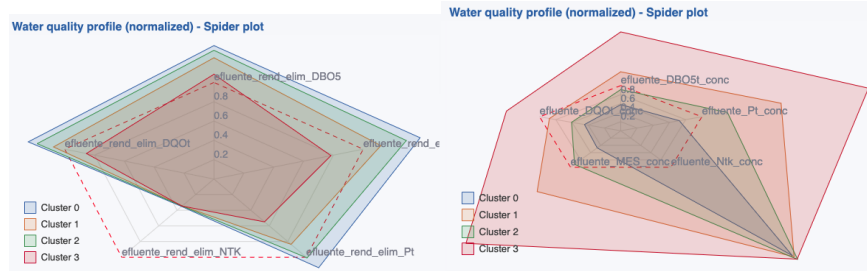


(a) Monitoring configuration parameters.



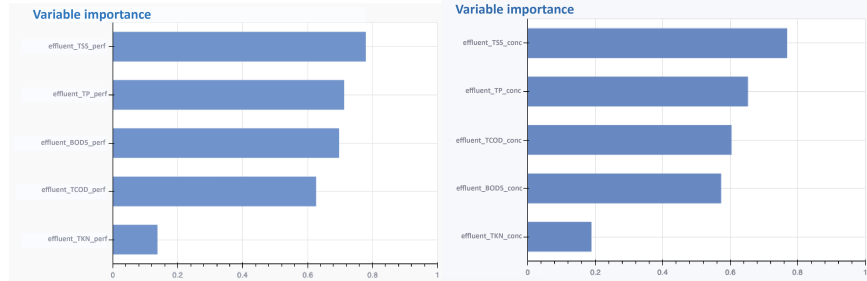
(b) Water Quality Line Chart (Performance).

(c) Water Quality Line Chart (Absolute).



(d) Water Quality Spider Chart (Performance).

(e) Water Quality Spider Chart (Absolute).



(f) Water Quality Variable Importance (Performance).

(g) Water Quality Variable Importance (Absolute).

Fig. 3. Visual Analytics Water Quality Monitoring Platform.

The confusion matrix can visualize model's performance in Figure 6, which shows how many of the values predicted by the model were correct according to the labels (real data).

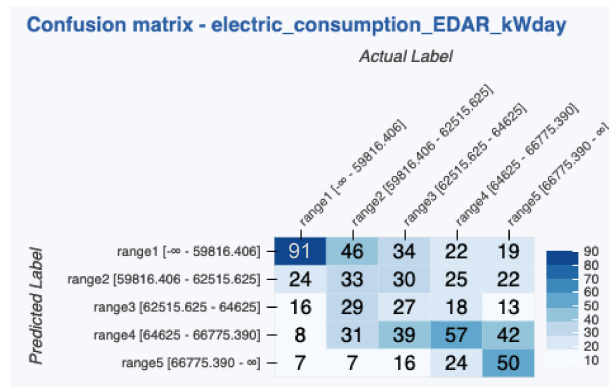


Fig. 6. Confusion matrix for the electric model.

In addition, the developed platform shows the relevance of the variables of the created model to the operator, as seen in Figure 7.

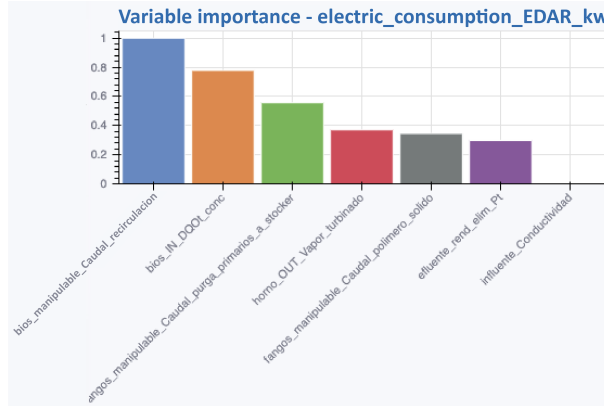


Fig. 7. Variable influence for the electric model.

Finally, the dashboard presents the decision tree created for a specific variable (model), as seen in Figure 8.

#### D. WWTP Model Optimization

This platform component is complementary to the simulator, where a target interval (range) is set for the variable being modeled, and restrictions are placed on the variables that influence it. Once this has been done, optimal values can be obtained for each influential variable to guarantee the modeled variable's target with the given restrictions. For example, Figure 9 shows which values of the chemical concentrations must be used to obtain the lowest possible range of energy consumption for the WWTP.

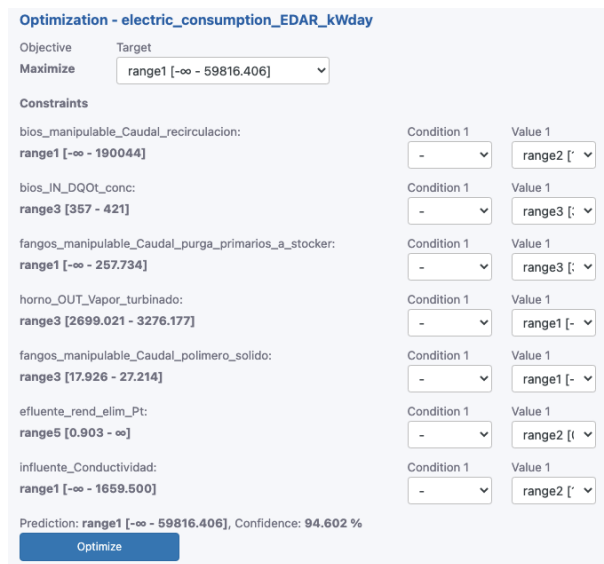


Fig. 9. Energy consumption model optimization.

## V. DISCUSSION

The end-user validated the operational improvement provided by the developed tools. This improvement comprises the following aspects concerning to their existing tools:

- **Observability:** it allows monitoring of the state of water quality through a visualization based on clustering.
- **Predictability:** operators can forecast how their WWTP will go.
- **Risk-free evaluation:** operators can validate how their system will perform if specific parameters change through simulation and optimization. This represents an essential advantage because, currently, they were required to test their actual WWTP, which could lead to damage if their operating variables were not correctly manipulated.
- **Interpretability:** The decision trees and variable importance graphs help the operators better understand their WWTP behavior.

The end user concluded that adequately trained and skilled staff could obtain the above benefits. Although initially this aspect might be interpreted as limiting, in the sense that if the plant management staff does not have the appropriate training, getting the benefits from the developed tools could be a complex, time-consuming and complicated task, in the end, it is considered as a positive situation by the end users as continuous education and training are part of worker's rights and company's obligation. Therefore, it is not seen as a limitation but as an opportunity to advance in innovation and continuous improvement.

Thus, incorporating new 4IR technology is suitable for the company and its operators, and the economic benefits from implementing the improvements that the user can identify through these tools are clear.

Finally, in addition to this qualitative and general validation, a quantitative validation of the performance of the developed Machine Learning models could be performed by the end user: On the one hand, Figure 10 shows the Confusion Matrix of the Water Quality Model, which is a tool for validating the model's predictive performance.

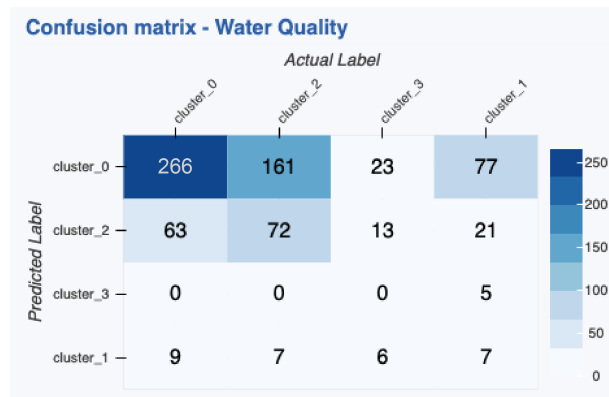


Fig. 10. Confusion matrix for water quality model.

On the other hand, the predictor importance graph is shown in Figure 11, which gives very valuable information about the

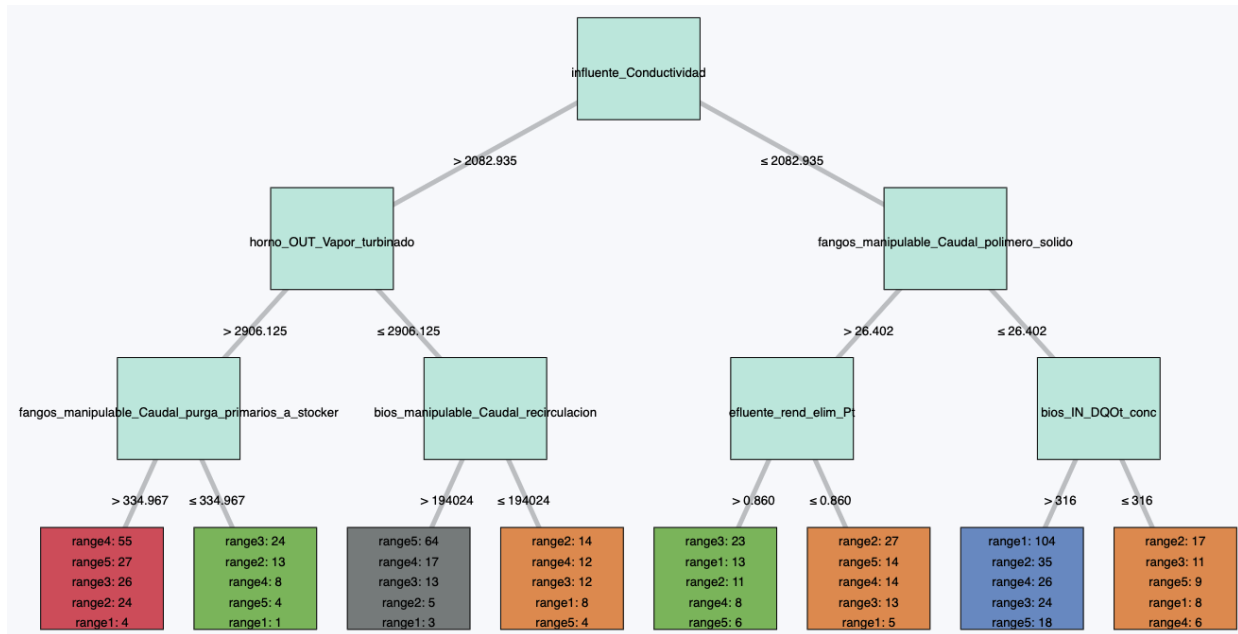


Fig. 8. Decision tree for the electric model.

variables that, according to the models constructed, have the greatest influence on the operation of the WWTP; the end user has confirmed these variables as those which greatly influence on the quality of the effluent water, which is the best proof of validation of the obtained results.

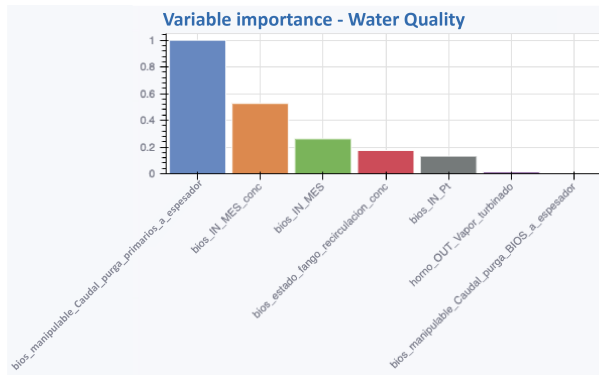


Fig. 11. Variable importance for water quality model.

## VI. CONCLUSION

This paper presents a visual analytics-based platform for WWTPs, called EDAR 4.0. Intuitive visualizations have great potential for supporting decision-making during the operation and management of WWTPs. The proposed tool allows users to identify relationships between key process variables through advanced data inspection. The developed tool allows WWTP operators to perform simulations and optimizations without risking real operation. The tool has been validated by WWTP domain experts, thus demonstrating its great potential as a

very valuable source of information for the day-to-day and long-term decision-making in WWTPs.

As future work for consolidating the use of the developed tools for the management of WWTPs, several possibilities are foreseen: (i) firstly, it is proposed to scale up the tool for a multi-plant implementation approach; (ii) secondly, the development of a dynamic ammonium controller through the scenario analysis and optimization functionalities provided by the developed tools is proposed, which would be an important novelty for WWTPs; (iii) thirdly, it is also proposed to carry out an in-depth study concerning usability; (iv) finally, the use of open-source Python libraries instead of RapidMiner (commercial software) is proposed for data analysis and model construction tasks to reduce costs and improve scalability.

## ACKNOWLEDGMENT

The authors would like to thank the Basque Government for supporting project “EDAR 4.0”, project partners *Mapner*, *Guascor Power*, *MaserMic*, *LKS Consultoría Tecnológica*, *Ibermática*, *IK4-Tekniker*, *Guascor Power I+D* and *Instituto Ibermática de Innovación (i3B)* for their contribution and WWTP operators *Utezea* and *Consorcio de Aguas de Gipuzkoa* for their collaboration.

The authors from Universidad EAFIT would also like to thank Vicerectoría de Descubrimiento y Creación from Universidad EAFIT for partially supporting this research.

## REFERENCES

- [1] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, *Visual Analytics: Definition, Process, and Challenges*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 154–175. [Online]. Available: [https://doi.org/10.1007/978-3-540-70956-5\\_7](https://doi.org/10.1007/978-3-540-70956-5_7)

- [2] A. Diez-Olivan, J. Del Ser, D. Galar, and B. Sierra, "Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0," *Information Fusion*, vol. 50, pp. 92–111, 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1566253518304706>
- [3] K. B. Newhart, R. W. Holloway, A. S. Hering, and T. Y. Cath, "Data-driven performance analyses of wastewater treatment plants: A review," *Water research*, vol. 157, pp. 498–513, 2019.
- [4] M. Maiza, J. Odriozola, A. Gil, G. Naveran, R. Basagoiti, I. Lecuona, U. Zurutuza, G. Urchegi, and A. Mañas, "Visual analytics for supporting the management of wwtps," in *Proceedings of the Young Water Professionals (YWP) conference*, 2017. [Online]. Available: <https://www.ywp-spain.es/event/congreso-ywp-2017/>
- [5] K. Jongrack, Y. Kwangtae, P. Wenhua, and K. Yejin, "Modified newton-raphson method to minimize calculation time for wastewater treatment plant simulation," *J. Korean Soc. Hazard Mitig.*, vol. 18, no. 5, pp. 319–326, 2018. [Online]. Available: <http://j-kosham.or.kr/journal/view.php?number=7861>
- [6] X. Flores-Alsina, C. Kazadi Mbamba, K. Solon, D. Vrecko, S. Tait, D. J. Batstone, U. Jeppsson, and K. V. Gernaey, "A plant-wide aqueous phase chemistry module describing ph variations and ion speciation/pairing in wastewater treatment process models," *Water Research*, vol. 85, pp. 255–265, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0043135415301160>
- [7] U. Jeppsson, C. Rosen, J. Alex, J. Copp, K. V. Gernaey, M.-N. Pons, and P. A. Vanrolleghem, "Towards a benchmark simulation model for plant-wide control strategy performance evaluation of wwtps," *Water Science and Technology*, vol. 53, no. 1, pp. 287–295, Jan 2006. [Online]. Available: <https://doi.org/10.2166/wst.2006.031>
- [8] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, "Fourier neural operator for parametric partial differential equations," 2020. [Online]. Available: <https://arxiv.org/abs/2010.08895>
- [9] K. A. Cook and J. J. Thomas, "Illuminating the path: The research and development agenda for visual analytics," Pacific Northwest National Lab.(PNNL), Richland, WA (United States), Tech. Rep., 2005.
- [10] J. van Wijk, "The value of visualization," in *VIS 05. IEEE Visualization, 2005.*, 2005, pp. 79–86.
- [11] J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia, and S. Liu, "A survey of visual analytics techniques for machine learning," *Computational Visual Media*, pp. 1–34, 2020.
- [12] J. K. Li and K.-L. Ma, "P6: A declarative language for integrating machine learning in visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, 2020.
- [13] A. Nawaz, A. S. Arora, W. Ali, N. Saxena, M. S. Khan, C. M. Yun, and M. Lee, "Intelligent human-machine interface: An agile operation and decision support for an ammox sbr system at a pilot-scale wastewater treatment plant," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 9, pp. 6224–6232, 2022.
- [14] A. Endert, W. Ribarsky, C. Turkay, B. W. Wong, I. Nabney, I. D. Blanco, and F. Rossi, "The state of the art in integrating machine learning into visual analytics," in *Computer Graphics Forum*, vol. 36, no. 8. Wiley Online Library, 2017, pp. 458–486.
- [15] D. A. Keim, T. Munzner, F. Rossi, and M. Verleysen, "Bridging information visualization with machine learning (dagstuhl seminar 15101)," in *Dagstuhl reports*, vol. 5, no. 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.
- [16] S. Liu, X. Wang, M. Liu, and J. Zhu, "Towards better analysis of machine learning models: A visual analytics perspective," *Visual Informatics*, vol. 1, no. 1, pp. 48–56, 2017.
- [17] C. D. Stolper, A. Perer, and D. Gotz, "Progressive visual analytics: User-driven visual exploration of in-progress analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1653–1662, 2014.
- [18] D. Jonker, R. Brath, and S. Langevin, "Industry-driven visual analytics for understanding financial timeseries models," in *2019 23rd International Conference Information Visualisation (IV)*. IEEE, 2019, pp. 210–215.
- [19] D. Sun, R. Huang, Y. Chen, Y. Wang, J. Zeng, M. Yuan, T.-C. Pong, and H. Qu, "Planningvis: A visual analytics approach to production planning in smart factories," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 579–589, 2019.
- [20] W. Wu, Y. Zheng, K. Chen, X. Wang, and N. Cao, "A visual analytics approach for equipment condition monitoring in smart factories of process industry," in *2018 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2018, pp. 140–149.
- [21] E. S. Matrosov, I. Huskova, J. R. Kasprzyk, J. J. Harou, C. Lambert, and P. M. Reed, "Many-objective optimization and visual analytics reveal key trade-offs for london's water supply," *Journal of Hydrology*, vol. 531, pp. 1040–1053, 2015.
- [22] M. Kim, Y. Kim, H. Kim, W. Piao, and C. Kim, "Operator decision support system for integrated wastewater management including wastewater treatment plants and receiving water bodies," *Environ Sci Pollut Res Int*, vol. 23, no. 11, pp. 10785–10798, Jun 2016.
- [23] S. Heo, K. Nam, J. Loy-Benitez, and C. Yoo, "Data-driven hybrid model for forecasting wastewater influent loads based on multimodal and ensemble deep learning," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 10, pp. 6925–6934, 2021.
- [24] W. Piao, C. Kim, S. Cho, H. Kim, M. Kim, and Y. Kim, "Development of a protocol to optimize electric power consumption and life cycle environmental impacts for operation of wastewater treatment plant," *Environ Sci Pollut Res Int*, vol. 23, no. 24, pp. 25451–25466, Dec 2016.
- [25] G. A. Avruskin, G. M. Jacquez, J. R. Meliker, M. J. Slotnick, A. M. Kaufmann, and J. O. Nriagu, "Visualization and exploratory analysis of epidemiologic data using a novel space time information system," *International Journal of Health Geographics*, vol. 3, no. 1, p. 26, 2004.
- [26] M. Anderberg, *Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*, ser. Probability and mathematical statistics. Elsevier Science, 2014. [Online]. Available: <https://books.google.com.co/books?id=7YTtBQAAQBAJ>
- [27] R. Haggarty, C. Miller, E. Scott, F. Wyllie, and M. Smith, "Functional clustering of water quality data in scotland," *Environmetrics*, vol. 23, no. 8 pp. 685–695, 2012. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2185>
- [28] H. Wong and B. Hu, "Application of interval clustering approach to water quality evaluation," *Journal of Hydrology*, vol. 491, pp. 1–12, 2013. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0022169413002096>
- [29] T. Vo-Van, A. Nguyen-Hai, M. V. Tat-Hong, and T. Nguyen-Trang, "A new clustering algorithm and its application in assessing the quality of underground water," *Scientific Programming*, vol. 2020, p. 6458576, Mar 2020. [Online]. Available: <https://doi.org/10.1155/2020/6458576>
- [30] M. J. Saary, "Radar plots: a useful way for presenting multivariate health care data," *Journal of Clinical Epidemiology*, vol. 61, no. 4, pp. 311–317, 2008. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0895435607003320>



**David Velásquez** received a B.S. degree in Mechatronics Engineering from Escuela de Ingeniería de Antioquia (EIA) in 2011. He got a master's degree in engineering at Universidad EAFIT, emphasizing Technical Systems Integrated Design in 2014. He works as an Assistant Professor in the Department of Systems and Informatics Engineering and as a researcher of the TICs Development and Innovation Research Group (GIDITIC) and the Design Engineering Research Group (GRID) at Universidad EAFIT.

His research interests include adaptive systems control design, mechatronics design, Industry 4.0, Machine Learning, Computer Vision, electronics optimization, Embedded Systems, Internet of Things implementation, and biomedical signal processing applications. He is also doing a Ph.D. in Informatics at the University of the Basque Country (Spain) in collaboration with research projects from the VICOMTECH research center.



**Paola Vallejo** received a B.S. degree in Systems Engineer from Universidad EAFIT in 2012. She got her Master's degree in Human-Computer Centered Systems at École Nationale d'Ingénieurs de Brest in 2012. She received the Ph. D. degree in Computer Science from Université de Bretagne Occidentale in 2015. Paola works as a Professor at the Department of Systems and Informatics Engineering and as a Researcher associated with the I+D+i on Information Technologies and Communications

Research Group at Universidad EAFIT. In particular, she is interested in model-driven engineering and human-computer interactions.



**Mauricio Toro** received a B.S. degree in Computer Science and Engineering from Pontificia Universidad Javeriana, Colombia, in 2009. Mauricio got a Ph.D. degree in Computer Science from Université de Bordeaux, France, emphasizing on Artificial Intelligence, in 2012. Mauricio was a postdoctoral fellow at the Computer-Science department at the University of Cyprus in 2013. Since 2014, Mauricio works as an Assistant Professor at the Department of Systems and Informatics Engineering and as a

researcher of the TICs Development and Innovation Research Group (GIDITIC) at Universidad EAFIT. His research interests include Artificial Intelligence, Industry 4.0, Machine Learning, Computer Vision, and Agricultural applications.



**Juan Odriozola** completed his studies in Automatic Engineering and Industrial Electronics at the University of Tecnun in 2010. In 2015, he obtained a doctorate degree for his thesis "Model based development of advanced mathematical tools for optimizing the operation of WWTPs", carried out at the University of Tecnun (Spain). In 2009-2015 he worked as a Junior Researcher at the Environmental Engineering Department of the Technological Centre CEIT, focusing his research on the model-based optimization of the design and operation of wastewater treatment plants (WWTP).

Since 2016 he works as a Researcher at the Data Intelligence for Energy and Industrial Processes Department of the Technological Centre VICOMTECH. His research interests simulation and optimization algorithms for modelling industrial processes and their integration with artificial intelligence techniques.



**Aitor Moreno** is a Computer Engineer from the University of Deusto in 1995, and a PhD in Artificial Intelligence from the University of the Basque Country. He is currently the head of the Department of Artificial Intelligence and Quantum Computing at Ibermática. In addition, he is a professor of the BigData Program at the University of Deusto, a professor of the Artificial Intelligence Program at the University of Navarra and a professor of Automation and Control at the University of the Basque Country, among

others. He manages projects related to the implementation of systems based on quantum optimization and simulation in hybrid systems, mainly focused on the extraction and automatic inference of relationships and knowledge in the form of semantic graphs. He participates in the management of European and national R&D projects in areas of application of Artificial Intelligence and Quantum Computing, and is a regular at international conferences, conferences, scientific publications and dissemination.



**Gorka Naveran** was born in Bilbao, Spain. He studied physics and electronic engineering specialising in control and regulation at the Universidad Complutense de Madrid. Has experience in the execution of singular projects (S.A.T.E. Malaga Airport) and more than 10 years linked to the management of energy efficiency, development of equipment and R+D projects.



**Mikel Maiza** completed his studies in Automatic Engineering and Industrial Electronics at the University of Mondragon in 2000. In 2003, he obtained a Doctorate degree with emphasis on Parallel Computing for Real-Time Systems, carried out at the University of York (United Kingdom). Since 2016 he works as a Senior Researcher at the Data Intelligence for Energy and Industrial Processes Department of the Technological Centre VICOMTECH. He has also been an External Professor at the University of Mondragon in 2002-2004, Associate Professor at the School of Engineering of the University of Navarra in 2009-2013 and Associate Professor at the Department of Applied Mathematics of the University of the Basque Country in 2015-2017. Currently, he collaborates as an External Professor with the Ecole Supérieure des Technologies Industrielles Avancées (ESTIA) since 2017. His research interests include parallel processing systems, heuristic algorithms and stochastic techniques of mathematical optimization and their integration with artificial intelligence techniques, for building stochastic models aimed at the simulation and optimization of processes and systems, as well as applications such as data mining, pattern recognition, automatic learning or early fault detection.

Since 2016 he works as a Senior Researcher at the Data Intelligence for Energy and Industrial Processes Department of the Technological Centre VICOMTECH. He has also been an External Professor at the University of Mondragon in 2002-2004, Associate Professor at the School of Engineering of the University of Navarra in 2009-2013 and Associate Professor at the Department of Applied Mathematics of the University of the Basque Country in 2015-2017. Currently, he collaborates as an External Professor with the Ecole Supérieure des Technologies Industrielles Avancées (ESTIA) since 2017. His research interests include parallel processing systems, heuristic algorithms and stochastic techniques of mathematical optimization and their integration with artificial intelligence techniques, for building stochastic models aimed at the simulation and optimization of processes and systems, as well as applications such as data mining, pattern recognition, automatic learning or early fault detection.



**Basilio Sierra** is Full Professor with the Computer Sciences and Artificial Intelligence Department, University of the Basque Country (UPV/EHU). He is the Co-Director of the Robotics and Autonomous Systems Group, RSAIT. He is also a Researcher in the fields of robotics and machine learning, where he is working on the use of different paradigms to improve robot's behaviours. He works as well in multidisciplinary applications of Machine Learning paradigms, in agriculture, natural language

processing, medicine, and so forth. He has published more than 50 journal articles, and several book chapters and conference papers.