Management and Industrial Engineering

Diego Carou
Antonio Sartal
J. Paulo Davim *Editors*

Machine Learning and Artificial Intelligence with Industrial Applications

From Big Data to Small Data

Springer

Management and Industrial Engineering

Series Editor

J. Paulo Davim, Department of Mechanical Engineering, University of Aveiro, Aveiro, Portugal

This series fosters information exchange and discussion on management and industrial engineering and related aspects, namely global management, organizational development and change, strategic management, lean production, performance management, production management, quality engineering, maintenance management, productivity improvement, materials management, human resource management, workforce behavior, innovation and change, technological and organizational flexibility, self-directed work teams, knowledge management, organizational learning, learning organizations, entrepreneurship, sustainable management, etc. The series provides discussion and the exchange of information on principles, strategies, models, techniques, methodologies and applications of management and industrial engineering in the field of the different types of organizational activities. It aims to communicate the latest developments and thinking in what concerns the latest research activity relating to new organizational challenges and changes world-wide. Contributions to this book series are welcome on all subjects related with management and industrial engineering. To submit a proposal or request further information, please contact Professor J. Paulo Davim, Book Series Editor, pdavim@ua.pt

More information about this series at https://link.springer.com/bookseries/11690

Diego Carou · Antonio Sartal · J. Paulo Davim
Editors

Machine Learning and Artificial Intelligence with Industrial Applications

From Big Data to Small Data

Springer

Editors
Diego Carou
School of Aeronautics and Space Engineering
Universidade de Vigo
Ourense, Spain

Antonio Sartal
Faculty of Economics and Business Administration
Universidade de Vigo
Vigo, Spain

J. Paulo Davim
Department of Mechanical Engineering
University of Aveiro
Aveiro, Portugal

ISSN 2365-0532 ISSN 2365-0540 (electronic)
Management and Industrial Engineering
ISBN 978-3-030-91005-1 ISBN 978-3-030-91006-8 (eBook)
https://doi.org/10.1007/978-3-030-91006-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

In the last few decades, data have become a key asset, not only for industry but also for society itself. In this context, the concept of big data has emerged, becoming extremely popular and trendy. When it comes to industry, companies must make the most of the available data to improve their performance and maintain competitiveness. Thus, artificial intelligence and, particularly, machine learning strategies are being extensively researched and implemented in different industrial sectors for multiple applications.

Although it is clear that some large companies manage data that can be categorized as big data, other ones are less demanding when it comes to data processing. However, these companies may also find important benefits from the same techniques used by the larger ones. In this regard, the present book was conceived with these two types of companies in mind. Thus, the book aims to be a source of insight for companies in different industrial sectors, whatever their size.

The objective of this book is to provide readers with a comprehensive work on machine learning and artificial intelligence. It is arranged in two parts. The first one presents the general context of the book and is composed of two introductory chapters. The second part is devoted to the presentation of case studies in various types of industrial activities. Thus, it comprises seven chapters. The contributions were written by researchers and scientists from academia, industry, and R&D centers, offering readers different enriching approaches.

Finally, we would like to thank all book contributors for their efforts and valuable contributions, as well as Springer for its support throughout this process.

Ourense, Spain — Diego Carou
Vigo, Spain — Antonio Sartal
Aveiro, Portugal — J. Paulo Davim

Contents

About the Editors

Diego Carou is an assistant professor at the University of Vigo. He received his Ph.D. degree in industrial engineering from the National University of Distance Education (UNED) in 2013. He has international postdoctoral experience in the manufacturing process at several European universities. His interests include Industry 4.0, manufacturing and sustainability.

Antonio Sartal is a postdoc researcher in the Department of Business Management and Marketing at the University of Vigo, Spain. He managed the Department of R&D of a food multinational for the past 10 years, until he joined REDE, a multidisciplinary research group working on technology management and organizational innovation. He received his Ph.D. degree in business management from the University of Girona (UdG) in 2015. His research interests include the intersection of lean thinking, innovation management, decarbonization and Industry 4.0 technologies.

J. Paulo Davim is a full professor at the University of Aveiro, Portugal. He is also distinguished as honorary professor in several universities/colleges in China, India and Spain. He has more than 30 years of teaching and research experience in Manufacturing, Materials, Mechanical and Industrial Engineering, with special emphasis in Machining & Tribology. He has also an interest in Management, Engineering Education and Higher Education for Sustainability.

A Note on Big Data and Value Creation

Miguel Angel Moreno-Mateos and Diego Carou

Abstract In the last years, big data has been increasing its popularity not only in the academic and industrial fields but also among the general public. In this regard, a bunch of state-of-the-art applications are arising, e.g., autonomous driving, crime forecasting techniques, medical diagnosis and smart cities. All these have a common denominator: handling large sets of data. However, huge amounts of data do not represent any value just by themselves, but they require further analysis. For this reason, advanced analytics is used to extract useful information and create value from raw data. Here, machine-based methods are proving to be adequate solutions in order to analyze data in different fields. The present chapter aims at providing an introduction to big data with the focus on value creation. To this end, it first reviews the origin and main features of big data to deliver a general context. Then, the pipeline to create value from raw data is explained. Its multiple steps, i.e., data generation, acquisition, storage, analytics, visualization and value creation, comprise techniques from diverse areas of knowledge. Finally, to illustrate how big data is currently being applied, we present a bunch of remarkable applications in the fields of bioengineering and medicine, economy, environment, industry and society.

M. A. Moreno-Mateos (✉)
Department of Continuum Mechanics and Structural Analysis, University Carlos III of Madrid, Avda. de la Universidad 30, 28911 Leganés, Madrid, Spain
e-mail: migmoren@ing.uc3m.es

D. Carou
Departamento de Deseño Na Enxeñaría, Universidade de Vigo, Campus As Lagoas, sn, Ourense, Spain

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8_1

1 Introduction

Even though in the last decades there has been an exponential increase of the share of data in the economy and society, data and analysis are almost as old as men. Already back in 3000 BCE in Mesopotamia, the Sumerians were identified as able to store, manipulate and communicate information (Termanini, 2020). Humans found several methods throughout history for preserving knowledge such as the clay table, codex, concertina, scroll and, later, books, which became the standard for centuries thanks to the development of the printing press (Liesaputra & Witten, 2012).

The book is one of the most iconic elements of humanity. They help in creating and expanding information and knowledge throughout the world. Also their success raised concerns as the growth in publication rates would finally result in difficulties to store and manage libraries as, for instance, identified Fremont Rider (Lucker, 1994) and Keyes Metcalf (Jifa & Lingling, 2014). Paper books started to get substituted, or complemented, with electronic books. These are mainly the same but the electronic version comprises novel functionalities, as reported by Liesaputra and Witten (2012). The electronic book allows easy accessing, marking, searching and connecting the content to other sources of information (e.g., a dictionary).

As population grows, undoubtedly information increases and its management becomes more challenging. Notable efforts were taken in the past to organize the information, for example, the *Encyclopédie* (Zhirov et al., 2010). As times went by, new solutions were required to properly manage the information. These solutions were only possible with technological developments that enabled new platforms, such as the electronic counterpart of the *Encyclopédie*, i.e., the Wikipedia.

Technology plays a crucial role in the capacity of not only maintaining and organizing large amounts of data but also analyzing the data and helping in the decision-making process. The importance of technological development for processing data is clear when attending to some examples. Back in the 1940s, one of the most important scientific endeavours was the Manhattan project. The project was intended for the development of nuclear weapons and to process data (e.g., for calculations). To this end, the teams used analogue IBM computers that were state-of-the-art at that moment (Atomic Heritage Foundation, 2014; Feynman, 1992). During the Second World War, Alan Turing invented an electro-mechanical machine, the Bombe, to decipher secret messages from the German enemy, who were using the Enigma machine. His device was able to solve mathematical problems that were described by algorithms. In the Apollo project, a digital computer was used as guidance. The computer had a weight of 65 pounds and an erasable memory with a capacity of 2,048 words (Erickson, 2021; MIT, 1969). Decades later, today the Tesla computer solution for the autopilot car has about 250 million gates or 6 billion transistors and includes 12 A72 CPUs (Talpes et al., 2020). Overall, computing plays a major role in the advance of big data and data analytics. The examples previously highlighted can be associated with the three computing eras that show the progress in the field: tabulating (1900s–1940s), programming (1950s–present) and cognitive (2011–) (Gupta et al., 2018).

The amount of data generated is increasing at enormous rates, especially in the last few years. To illustrate this, not so far ago in 2010, Eric Schmidt, former Google CEO, stated that, by that time, the amount of data created in two days was equal to that created from the dawn of civilization until 2003 (Gupta et al., 2018). The numbers are spectacular, hence some interesting estimations can be highlighted. Later, in 2020, an estimation by the International Data Corporation (IDC) indicated that 64.2 zettabytes were created or replicated worldwide (IDC, 2021), which is equivalent to a pile of books from the Sun to Pluto several times.

Currently, data are at the centre of most, if not all, types of private and public activities. In fact, driven by major American tech companies (e.g., Alphabet (Google), Amazon, Apple, Cloudera, Facebook, IBM, Microsoft, Oracle, Salesforce and Teradata) and their Chinese counterparts (e.g., Alibaba, Baidu, Xiaomi and Tencent), the concept of data economy has emerged and occupied an important place (Aalst et al., 2019; Lammi and Pantzar, 2019). When attending to the big data market, Adroit Market Research (2020) forecasts its size at $267 billion by 2025.

In the last few years, a large number of terms and concepts around data and data analysis have been developed. Moreover, of great importance is the influence that the COVID-19 pandemic has on the digitalization of society, boosting internet usage, videoconferencing and entertainment streaming services (WEF, 2021). This fact makes one wonder how the shape and structure of data look like and what the best strategies to analyze these data are. In order to help the reader understand this matter, the present chapter offers a comprehensive introduction to the topic. The second section is intended for clarifying the main ideas related to big data by way of context. The origin of big data is addressed and its main features are described, also including a comparison with small data. The third section focuses on value creation and the introduction of real examples of applications in different areas. For this purpose, the general pipeline for creating value from raw big data is first depicted, paying special attention to the steps involved, i.e., data acquisition, data storage, data analytics and visualization. Then, we have elaborated on some interesting applications by a state of the art revision, namely medical forecasting and theoretical biology (e.g., cancer diagnosis and tissue histology), supply chain management and industry 4.0., smart cities, insurance services, crime forecasting, high-velocity trading algorithms and environmental preservation.

2 Understanding Big Data

Obviously, big data is an extension of the data term. Although the reader may be familiar with the concept, it is interesting to present a definition that helps position data in the context of this chapter. So, the Cambridge dictionary (https://dictionary.cambridge.org) defines data as: "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer". The interest of this definition relies on the fact that it connects data to computers and, finally, to decision-making processes. This is what this chapter deals with: the storage and management of (big) data to lastly analyze the information with advanced analytics, helping users in the decision-making.

Big data is a term that was first used in 1997. The NASA (National Aeronautics and Space Administration) scientists Michael Cox and David Ellsworth coined the concept in the "Application-controlled demand paging for out-of-core visualization" paper presented at an IEEE conference (Cox & Ellsworth, 1997; Wang et al., 2018; Bydon et al., 2020). But, what exactly is big data? Unfortunately, there is not a single answer to this question. Conventionally, big data has been related to high volumes of data (Faraway & Augustin, 2018). For instance, Zhong et al. (2016) stated that the term refers to data in the exabytes range, while Shukla et al. (2020) indicated that datasets of sizes in the petabytes range are also being used. This results in datasets composed of billions to trillions of archives (Altaf-Ul-Amin et al., 2014). However, big data can also have different meanings and relate to a wide variety of the data or a high velocity for data gathering (Gao et al., 2020). In addition, big data can also be understood from a wider perspective (Faraway & Augustin, 2018). Thus, big data may also refer to the extent, impact and mindshare of the phenomenon.

To understand the complexity and nature of data, several models developed in the past can be introduced. For instance, the 3 V's model has been largely discussed in the later years. New models are also emerging to provide a bigger picture of what big data is. These recent approaches enlarge the former one by adding additional V's such as the 4 and 5 V's models (Liu et al., 2016), and even larger approaches are being proposed (e.g., 7 V's and 51 V's) (Leung, 2021; Khan et al., 2019). The 5 V's of big data is one of the most popular models and defines five main features of big data: volume, velocity, variety, veracity and value (Ishwarappa, 2015). First, big data, mostly unstructured, is massive in volume. For instance, each minute hundreds of hours of video are uploaded on Youtube and over 200 million emails are sent through Gmail. Second, data flows are extremely fast (i.e., speed in data creation, processing, storage and analysis) and information can quickly become outdated. Third, data come from a wide variety of sources (Azeem et al., 2021). The big data typology according to the data sources was classified by Calza et al. (2020) in the categories of social media, machine-generated data (i.e. M2M communication), sensing, transactions and Internet of Things (IoT). Also, the rise of biometric data with the growth of electronic devices entails new challenges (Ross et al., 2020). Such variety makes it more difficult to put data together in relational databases. Fourth, due to the high volume, velocity and diversity of data, it is a challenge to guarantee the veracity of data, as it usually may be incomplete or incorrect. Fifth, value is a crucial feature of big data, and identifying trends and patterns on it brings a competitive advantage to enterprises. This set of features is well defining of the paradigm shift brought along with big data.

As a counterpoint of big data, the concept of small data is also finding its place, for instance, in the business community. Small data refers to amounts of data limited in size, non-continuously collected, with narrow variety and intended to be used for

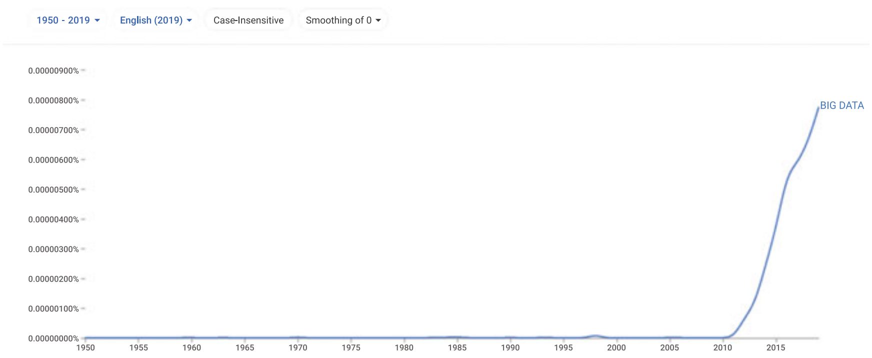


Fig. 1 Results of "big data" from 1950 to 2019 provided by Google Books Ngram Viewer (http://books.google.com/ngrams)

understanding rather than predicting (Faraway & Augustin, 2018; Kitchin & Lauriault, 2015). Kitchin and Lauriault (2015) performed a comparison between big data and small data. On the one hand, small data is limited in volume and depicts individual samples with reduced variety (e.g., a laboratory experiment with data acquisition). Its resolution is coarse, its velocity is low and it is not easily scalable. On the other hand, big data is massive in volume and covers entire heterogeneous populations. Its resolution is fine, its velocity is fast and it is easily scalable. Overall, big data is more flexible than small data.

As Witten et al. (2011) pointed out, there is an increasing gap between the "generation of data" and "our understanding of it". As the generation of data seems to be ever increasing without glimpsing a change in the trend, humans started to pay more attention to that and tried to solve all related problems by developing new methodologies and software solutions. Within this framework, the term big data started to get popular in the last years, especially in the 2010s (Liu et al., 2016). By way of example, Fig. 1 introduces a search of the term "big data" at Google Ngram Viewer from 1950 to 2019. The results reveal how the term has remarkably increased its popularity from the 2010s on (Michel et al., 2011).

3 From Data to Value Creation

For humans and companies data by itself has no value. According to Monino (2021), there is a connection between data, information and knowledge. Thus, data can be understood as "a collection of facts, observations, or even raw elements that are not organized or processed in any particular manner and that contain no discernible meaning"; Information "is an interpreted flow of data" and, finally, knowledge "is created and organized from this flow of information". Therefore, the value of data appears when it enables users in the decision-making or, in other words, when data are converted into knowledge. Jifa and Lingling (2014) further extended these concepts

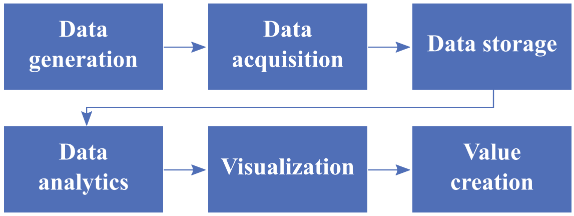


Fig. 2 Flow from data generation to value creation (based on Saggi and Jain (2018))

adding wisdom and depicting the set of these four terms (data, information, knowledge and wisdom) by means of a hierarchical pyramid. In this sense, wisdom is on top and can be understood as "an extrapolative and non-deterministic, non-probabilistic process. It calls upon all the previous levels of consciousness, and specifically upon special types of human programming (moral, ethical codes, etc.)."

The common steps required to create value from data are depicted in Fig. 2. Different sources, such as humans or machines, generate data. Typically, data are obtained from social networks, mobile devices, Radio Frequency Identification (RFID) technology, industrial sensors installed on industrial machinery and management information systems. Afterwards, these data are transferred to the data warehouse with proper filtering and cleaning. Then, they are stored performing activities such as data clustering, replication and indexing (Saggi & Jain, 2018). Data analytics or, when referring to big data, big data analytics, is the core of the value creation process. The concept of big data analytics is an etymological evolution of data mining and a part of business intelligence that refers to the set of methods that aim at finding patterns and trends in massive unstructured data. In addition, data analytics methods are diverse depending on the techniques on which they are based. Some of these techniques are machine learning (Cichos et al., 2020; Thyago et al., 2019; Zhang et al., 2020; Wang et al., 2018), artificial neural networks for prediction and classification (Jeswal & Chakraverty, 2019), deep learning (Gopakumar et al., 2018; Saxe et al., 2021), dynamic Bayesian networks for building graphical probabilistic models (Amin et al., 2021), text analytics, natural language processing (Wang et al., 2020), text mining to analyze unstructured text with Hidden Markov Models for speech, handwriting and gesture recognition (Agrawal, 2014), social network analysis, signal processing and optimization methods (Saggi & Jain, 2018; Özemre & Kabadurmus, 2020). Finally, visualization accounts for the strategies to display information in such a way that it can be easily understood (i.e., diagrams, excel sheets, histograms, maps, etc.). Overall, this pipeline aims at creating value as the output information is used in decision-making.

Big data requires a big infrastructure that supports all the steps shown in Fig. 2. In this concern, there is a rising trend to optimize computational resources by means of cloud computing. Cloud computing is based on virtualization to dynamically allocate physical computers' capacities (e.g., storage, CPU) to individual computing processes (Jamsa, 2012). In this way, hardware can be standardized, the most efficient use of resources can be done and better technical support is guaranteed by the service provider. Infrastructure as a service (IaaS), Platform as a service (PaaS) and

Software as a service (SaaS) clouds gather on-demand capacities obtained over a network (Sartal et al., 2020). Xinhua et al. (2013) reviewed Big data as a service (BDaaS), a variant of the former service infrastructures which focuses on big data. Google, with Google Drive, and Amazon, with Amazon Web Services, are great examples of service cloud providers. Even though firewalls protect information at cloud services, it is usual to find hybrid models where in-house IT capacities supports cloud computing. In this way, confidential information is kept within the private IT systems (Laudon, 2019). The infrastructure just depicted enables a framework for the steps involved in the value creation pipeline. In the following, these steps are to be covered with more detail.

Data acquisition is understood as a workflow that aims at gathering and cleaning data before storing it or, in other words, to make them ready for storage and further analysis purposes. The processes involved adjust to the big data features: high volume, high velocity and high variety. Also, little initial value (i.e., value is created at the end of the value creation pipeline) and unclear veracity are features of the raw data. The acquisition process starts with the storage of raw data in scalable storage solutions, e.g., Hadoop Distributed File System (HDFS) and NoSQL. Then, the data undergoes a curation and reorganization process via Structured Query Language (SQL)-based software. The process is supported by: (i) protocols (i.e., open protocols AMQP and Java Message Service), (ii) frameworks for data collection; and, (iii) proper enabling technologies (Lyko et al., 2016). Additional software tools that help in data acquisition are Storm, S4, Kafka and Flume. Braun et al. (2018) developed a case study about data generation and acquisition within basketball matches in the NBA in R programming language. The authors elaborated an analysis to choose the most convenient players for a match by tidy and light data files gathered from offensive and defensive actions observed over the course of the game. Regarding data acquisition, a further remark is that the way in which data is acquired determines the following step in the value creation, i.e., its storage.

Because of the increasing data generation, data storage is gaining importance in the whole value creation process. Data are generally stored by using data centres based on magnetic storage technology (Bhat, 2018). However, there is a need to improve the storage capabilities and, so, new technologies are being developed and analyzed as reviewed by Bhat (2018). Some of the technologies include Optical data storage (ODS), DNA data storage (DDS) and Holographic data storage (HDS). First, ODS, born with the Compact Disc in the 1980s, uses optical lithography to record information on a layer beneath the surface of the disc. The diffraction barrier limits the storage density to some terabytes per DVD-shaped disc, hence hindering its use for big data. However, recent photonic approaches suggest that this technique could reach large enough storage densities to be used in big data. Second, and based on DNA nucleotide structure, DDS was thought in the 1960s to be a promising high-density storage technology. This technique reproduces the way in which biological information is stored in DNA chains built by nucleotides (a double-helix-shaped structure). Three are the main advantages of this storage medium: it is a high-density storage medium where a bit can be recorded with a few atoms, it is volumetric instead of planar and it is quite stable in time. However, DDS encounters some problems, namely the difficulties in avoiding errors when synthesizing and sequencing the chain of nucleotides. Also, the fact that it is a time-consuming process. Finally, HDS defines the frames to record information in volume and not only on planar surfaces. By means of multiplexing, data are recorded as holograms allowing greater storage capacities. Challenges in HDS technique relate to the unavailability of recording and reading media. All in all, storage is necessary to handle big data, but useful information can only be obtained by using analysis and processing data techniques.

Once data have been properly gathered, they should be analyzed in order to extract knowledge. In this regard, traditionally mathematical models have been used to analyze samples of data (small data) under a model-driven approach. However, with the advance of big data, the paradigm should change to a data-driven one (Cheng et al., 2016). As a general premise, one could state that humans learn from experience, and so do machines (Lake et al., 2017). Here, experience denotes past information, namely digital data gathered for analysis, that the learner uses to make predictions (Mohri et al., 2012). Machines that are able to solve learning problems implement predictive machine learning algorithms. On these algorithms, typically as data-driven models, merge the notions of statistics, optimization and probability. An example of a learning problem is that of predictive maintenance, where data collection infrastructure (e.g., sensors) harvests data from machines and processes within the industrial environment in order to predict failure, hence action (Carvalho et al., 2019). A related concept to machine learning is that of intelligence. This latter, denoted as Artificial intelligence within a computing framework, is essentially linked to neuronal and synaptic activity.

Artificial neural networks (ANNs) are a pioneering tool in data analytics. They were born a few years after psychologist Donald Olding Hebb developed the theory of neural plasticity. He postulated that physiological changes take place in brain synaptic structures depending on the way neurons are activated. The value of the synapsis (i.e., the strength of the connection of two neurons) becomes higher if the neurons are activated simultaneously, and smaller if activated non-simultaneously. ANNs are based on this approach to dynamically adjust the weight of the synapsis according to the activation of artificial neurons. This is typically know as the training of the network. These features make ANNs useful decision-making algorithms for data analysis. Figure 3 depicts a general structure of a network with several layers, neurons and connections. For the sake of example, neuromorphic vision approaches human vision to efficiently sense, store and process visual information on human-made devices and on the fly. In this regard, Zhu et al. (2021) have developed a 1024-pixel optoelectronic sensor with carbon nanotubes and quantum dots as active materials and prove its sensitivity to light stimuli. A particular case of ANN are convolutional neural networks (CNN), widely used for image processing purposes. CNNs slide filters over the input images to better predict their outcome (Desai & Shah, 2021). To a more general extent, Yang et al. (2021) have recently proposed a deep-learning-based framework to design composites (i.e., design of its physical properties) via optimization of its geometrical hierarchy. The model comprises a conditional generative adversarial neural network (cGAN), which generates data from statistical analysis of the training set, to link the material's microstructure, the design and the

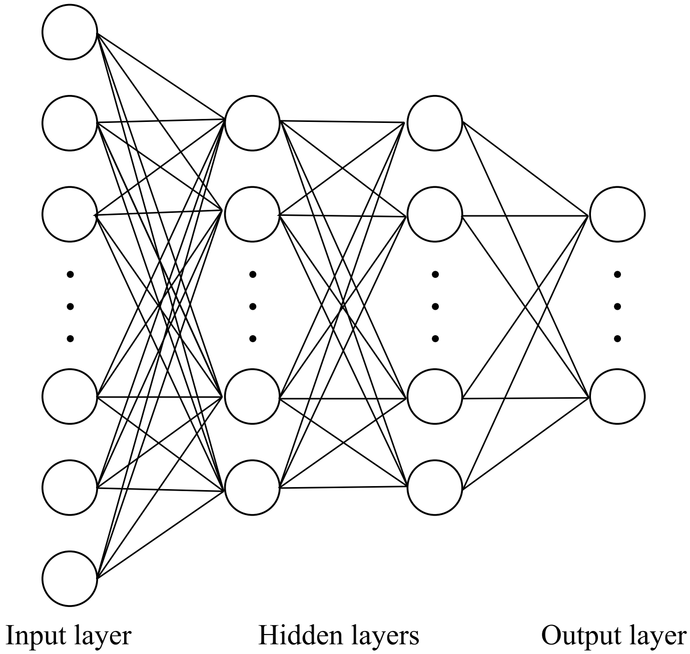


Fig. 3 Artificial Neural Network composed by an input layer, an output layer and two intermediate layers. The number of neurons of each layer is generalized to an arbitrary number

structural performance spaces. Further approaches for materials design are based in the optimisation of magneto-active composites via metaheuristic and machine learning-evolutionary algorithms (Pelteret et al., 2018; Sun et al., 2021). At this point, data are already processed, and it follows proper ways to display information, which enable the creation of value.

Visualization is a fundamental tool in the decision-making process. It consists of all the techniques to display output information in a useful and a suitable way to create value. Throughout history, many ways of displaying information have been used. According to Friendly (2008), some iconic milestones can be distinguished. Before the seventeenth century, early maps and diagrams appeared as geometric diagrams, sky maps and navigation maps. Claudius Ptolemy, in the second century, projected the spherical Earth into a longitude and latitude, and Frisius and Tartaglia, in the sixteenth century, developed triangulation methods to accurately determine locations from a map. Measurement of time, distance and space were the major concerns in the seventeenth century. This century is said to be where visual thinking started. Later, in the eighteenth century, new ways of data representation were developed, namely isolines, timelines, curve fitting and interpolation. William Playfair (1759–1823) is widely accepted to have invented most of the graphical representations used today. For example, in 1801, he created a pie-circle-line chart to represent taxes per capita in several nations.

In the first half of the nineteenth century, an abrupt growth of statistical charts took place (e.g., bar and pie charts, histograms, scatter plots, etc.). Also, thematic mapping experienced a vertiginous growth with atlases including weather, tides, economic social and medical information. These graphic charts evolved in the second half of the century, known as The Golden Age of statistical graphics (Friendly, 2008). Early in the twentieth century, some innovations were linked to the Second World War. Abraham Wald, an Hungarian mathematician, worked for the United States during the Second World War to find the optimal amount of armour for war planes (Atanasiu, 2021). After analyzing large amounts of data, he found that areas with fewer recorded shots required the most armour. He claimed that it was in these areas where the planes that did not return had the most dangerous shots. In the period from 1900 to 1950, stated by Friendly (2008) as the Modern Dark Ages, the enthusiasm for graphical data visualization slowed down and few graphical innovations were introduced. However, this situation changed from 1950 on, with the rebirth of data visualization and the rise of computing power and the personal computer.

From 1950 to present, increasing computing power has made it possible to perform multidimensional analysis and determine relations in higher dimension spaces (e.g., High-D software tool from the company macrofocus to interactively visualize multi-dimensional datasets). In 2018, scientists working with the Oak Ridge Leadership Computing Facility (OLCF) used Titan supercomputer, one of the most powerful computers in America, to elaborate a remarkable visualization tool called SIGHT. By means of this exploratory visualization tool, researchers can improve models even before performing the simulations. Scientists from the University of Virginia applied this kind of visualization to unveil the effects of laser ablation on the surface of metals. A piece of their results is shown in Fig. 4, which depicts the microscopic rendering of the nanoparticles generated on the surface during the ablation process. A special matter, when referring to big data, arises when the results of the analysis are presented as text or in tabular form. This makes it difficult for users to take full advantage of them, extract knowledge and get insightful information (Leung, 2021). In this

Fig. 4 Microscopic representation of the transformations that occur on metallic surfaces during laser-based processes: Left (Oak Ridge National Laboratory, 2019); Right (Oak Ridge National Laboratory, 2018). Courtesy of Oak Ridge National Laboratory, U.S. Dept. of Energy

respect, several visualization techniques were developed to help in getting knowledge from the data, such as the blockmap, circle packing, circular network diagram, heatmap, parallel coordinate, streamgraph, sunburst, treemap; and, more recently, for big data such as the libraries: Chart.js, D3.js, Dygraphs, FusionCharts and RHadoop; the platforms: Gephi, Power BI, R and Tableau; and, the services: CartoDB, Dundas Dashboard and ECharts (Kahil et al., 2020). Overall, data visualization has become more varied, interactive and mature than ever before, allowing the development of the field of visual analytics in parallel to the advance of computer science (Wang et al., 2019; Earnshaw, 2019). The next paragraphs address a bunch of state-of-the-art applications derived from the value creation framework.

In medicine and bioengineering sciences, pioneering work is being done for the benefit of human health. Cancerous tissue detection by means of noninvasive image processing offers a low-risk diagnostic path suitable to supplement traditional techniques (e.g., mammography). Here, Ali and Aittokallio (2019) have developed a methodology for implementing an optical biopsy, together with a neural network (the public CNN ResNet50), and utilize it to classify breast cancer tissues. Computer diagnosis, aided by machine learning, is also depicted by Ali and Aittokallio (2019) as a useful tool for predicting the personalized physiological response to oncological treatments. Big data based on Next Generation DNA Sequencing (NGS) databases and other repositories is paving the way towards more accurate drug response prediction from genome information.

In the work “Fifty Shades of the Brain”, Budday et al. (2020) presented a detailed review of the approaches for mathematically modelling the most important tissue in the human body: the brain tissue. Among other results, they open doors for linking with meaningful machine learning algorithms based on big data. Within this scope, Schroder et al. (2021) elaborated a predictive approach where machine learning is used to characterize traumatic brain injury after several prototypical accidents. Beyond being a medical predictive tool, this would enable accident reconstruction by analyzing the caused physiological damage. To a more general extent, Alber et al. (2019) reviewed the increasing significance of data (amount and value) in the biological, biomedical and behavioural sciences. The authors advocate the use of computational models, together with machine learning schemes, to improve their predictive capacity and scope. Overall, big data analytics is quite promising for value creation in the scope of human health.

Also in industry, big data analytics is remarkably increasing its role, particularly in the context of Industry 4.0, with not only a technological transformation, but also a business transformation. Real-time optimization and prediction of business flows is the major representation of these sweeping changes (Schmarzo, 2013). Here, information systems play a crucial role. Despite traditionally focusing on structured small data, rising interest in big data analytics is boosting the implementation of more sophisticated analysis methods. For instance, BIM BigSheets, developed on the Hadoop framework, supports companies to handle massive amounts of unstructured data from the web, providing the results directly through a web browser. Also, OLAP (Online Analytical Processing) cubes are a tool to perform multidimensional analysis with reduced computational power. This new industrial revolution is based on the vast generation of data gathered from multiple sources that, then, are analyzed by advanced (big) data analytics (Carou, 2021). As a result, companies are able to create value in a number of ways for them and for their stakeholders. Note that this value is bigger when the stakeholders are aligned on a common interest (Pesce et al., 2019). It is interesting to cite initiatives related to data such as the one by Rolls Royce, the R^2 Data Labs (Rolls Royce, 2021). The company aims at delivering more value to its customers by means of data innovation and advanced data analytics, resulting in improvements in operational performance.

The complexity of the current globalized economy makes big data of special interest for the management of supply chains, especially the large multinational ones of the larger sectors (Hallikas et al., 2021; Shamout, 2020). To a more general extent, Kumar et al. (2020) report beneficial effects on series-based and causal demand forecasting. For this purpose, an artificial neural network learns from current trends, marketing events data, promotions and other factors. Furthermore, analysis techniques could be implemented in marketing policies. This would require inferring data about decision-making cycles of customers, via online shopping applications, social media and IoT (Wang et al., 2019). Overall, industry can get a competitive advantage if it succeeds in establishing analytics strategies to make their processes more profitable.

Data analytics also finds application in everyday life in cities and communities, where massive amounts of data are generated. Smart cities are sustainability-focused infrastructures able to improve the efficiency in the use of public services. For example, smart car parking, where clustering and anomaly detection from raw data would help identify unusual events such as parking spots affected by external factors (Zheng et al., 2014). Also, insurance services benefit from data analytics, e.g., car insurance by predicting customer risk score and inferring risk reduction techniques (Longhi & Nanni, 2020). A general concern is that greater flows of data call for greater cybersecurity solutions. Car driving, for example, should never be unsafe due to an IT connectivity matter (Levi et al., 2018).

Criminals and criminal organizations do not adjust to the principles of randomness, but they often depict behavioural patterns, which could be predicted. Crime forecasting based on big data helps search the most probable criminal of a crime that has already happened and define the occurrence likelihood for a certain type of crime in a certain area (Kumar & Nagpal, 2019). To this end, parameters such as the crime type, date, location and suspects are the sources to build crime profiles and deploy resources in an effective manner. For instance, The New York City Police Department (NYPD) can visualize, in seconds, a suspect's photograph and its criminal history. Furthermore, Feng et al. (2019) infer top dates with lowest and highest crime occurrence probability in Chicago, Philadelphia and San Francisco. Hence, they suggest setting holidays on the calmest dates and the largest deployment of forces in summer.

High-velocity trading algorithms are conceived to perform financial operations as fast as possible (i.e., interchanging financial assets). The decision-making process usually takes place in less than one second according to the evolution of market variables and humans are excluded from the decision (Han & Li, 2018). These automatic systems require safety measures to ensure a proper operation (Bell, 2013). For example, the action taken should be sensitive to the input parameter, but ought to have enough inertia against sudden meaningless fluctuations, i.e., random (noise) alterations of the value of enterprise's shares (Han & Li, 2018).

Finally, some other remarkable applications of big data can be found. For instance, Freihaut and Göritz (2021) conducted research to infer trends about mental stress by means of the computer mouse (Freihaut and Göritz 2021). Big data analytics is also useful in order to predict potential environmental risks and wildfire responses (Thompson et al., 2019). It helps governments calculate response times when facing natural disasters. In this way, the United States Forest Service uses big data to predict potential risks for fauna and flora. Overall, ecology emerges as a major concern where a proper data treatment and arrangement helps tackle such endeavour (Reichman et al., 2011). Many other applications can be found elsewhere such as in autonomous driving and military warfare (Peeters et al., 2021).

All the previously depicted fields of application are nice approaches in the way of value creation. They all provide a positive impact on the economy (i.e. competitive advantage in industry and high velocity trading) and on social well-being and safety (i.e. autonomous driving, crime forecasting and medical response prediction). From raw data, and by means of analysis algorithms, Becker (2016) states that value is sorted in multiple forms: prediction, planning, simulation, visualization, comparison and control capacities, almost all of them supporting the decision-making logic.

4 Conclusions

Data is a pivotal element for human evolution. It underlies the creation of knowledge to a more efficient and with a greater quality of life society and environment. The present chapter aimed at providing an introduction to the concept of big data and related ones. To this end, first, a global framework for understanding big data was developed. Big data has been reviewed starting from its origin in NASA to its significance in decision-making processes. Volume, velocity, variety, veracity and value are the five characteristics that describe big data according to the 5 V's model and that distinguish it from small data. Second, some guidelines concerning value creation were depicted. Optical, DNA and holographic technologies are relevant technologies to store data. Afterwards, several strategies to extract valuable information from the stored data are introduced. Inspired by brain neural processes, artificial neural networks have been presented as a pioneering tool in data analytics. It followed the revision, from a historical perspective, of data visualization approaches. These are the ultimate link of the chain for human to make valuable decisions. Third, a set of current fields of applications of big data were depicted. Starting with the creation of value in the medical and bioengineering sciences with predictive tools for the response to treatments and mechanical injury, then passing by applications and current projects in autonomous driving, crime patterns forecasting, high-velocity trading algorithms, industry, smart cities and ending with some other curious applications, big data has been proved to be a quite powerful framework to enhance the economy and society well-being. As Seneca the Younger already believed, nature has given us the seeds of knowledge, not the knowledge itself. Consequently, it is our responsibility to properly build it for the benefit of the society and the environment.

References

Adroit Market Research (2020). Analysis, expansion plans, key growth drivers, and business opportunities: Adroit Market Research. Available at: https://www.globenewswire.com/news-release/2020/04/09/2014372/0/en/Big-Data-Market-to-grow-at-12-CAGR-to-reach-267-billion-by-2025-Global-Insights-on-Key-Trends-Size-Leading-Players-Investments-Analysis-Expansion-Plans-Key-Growth-Drivers-and-Busin.html

Agrawal, D. (2014). Analytics based decision making. *Journal of Indian Business Research, 6*(4), 332–340.

Alber, M., Buganza Tepole, A., Cannon, W.R., De S., Dura-Bernal S., Garikipati K., Karniadakis G., Lytton W.W., Perdikaris P., Petzold L., & Kuhl E. (2019). Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *NPJ Digital Medicine, 2*, 115.

Ali, M., & Aittokallio, T. (2019). Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophysical Reviews, 11*(1), 31–39.

Altaf-Ul-Amin, M., Afendi, F. M., Kiboi, S. K., & Kanaya, S. (2014). Systems biology in the context of big data and networks. *Biomed Research International* 428570.

Amin, M. T., Khan, F., Ahmed, S., & Imtiaz, S. (2021). A data-driven Bayesian network learning method for process fault diagnosis. *Process Safety and Environmental Protection*, 150110–150122.

Atanasiu, R. (2021). Hidden assumptions. In: *Critical thinking for managers. Management for professionals*. Springer, Cham

Atomic Heritage Foundation (2014). Computing and the Manhattan project. Available at: https://www.atomicheritage.org/history/computing-and-manhattan-project

Azeem, M., Haleem, A., Bahl, S., Javaid, M., Suman, R., & Nandan, D. (2022). Big data applications to take up major challenges across manufacturing industries: A brief review. *Materials Today: Proceedings, 49*(2), 339–348.

Becker, T. (2016). Big data usage. In: *New horizons for a data-driven economy* (pp. 143–165). Cham: Springer International Publishing.

Bell, H. A. (2013). High frequency trading: Do regulators need to control this tool of informationally efficient markets? Cato Institute Policy Analysis No. 731.

Bhat, W. A. (2018). Bridging data-capacity gap in big data storage. *Future Generation Computer Systems, 87*, 538–548.

Braun, M. T., Kuljanin, G., & DeShon, R. P. (2018). Special considerations for the acquisition and wrangling of big data. *Organizational Research Methods, 21*(3), 633–659.

Budday, S., Ovaert, T. C., Holzapfel, G. A., Steinmann, P., & Kuhl, E. (2020). Fifty shades of brain: A review on the mechanical testing and modeling of brain tissue. *Archives of Computational Methods in Engineering, 27*, 1187–1230.

Bydon, M., Schirmer, C. M., Oermann, E. K., Kitagawa, R. S., Pouratian, N., Davies, J., Sharan, A., & Chambless, L. B. (2020). Big data defined: A practical review for neurosurgeons. *World Neurosurgery, 133*, e842–e849.

Calza, F., Parmentola, A., & Tutore, I. (2020). Big data and natural environment. How does different data support different green strategies? *Sustainable Futures, 2*:100029.

Carou, D. (2021). *Aerospace and digitalization. A transformation through key industry 4.0 technologies*. Springer.

Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. P., Basto, J. P., & Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering, 137*, 106024.

Cheng, S., Zhang, Q., & Qin, Q. (2016). Big data analytics with swarm intelligence. *Industrial Management & Data Systems, 116*(4), 646–666.

Cichos, F., Gustavsson, K., Mehlig, B., & Volpe, G. (2020). Machine learning for active matter. *Nature Machine Intelligence 2*, 94–103.

Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. In: *Proceedings of the 8th IEEE conference on visualization.* Los Alamitos, CA: IEEE Computer Society Press.

Desai, M., & Shah, M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN). *Clinical eHealth, 4*, 1–11.

Earnshaw, R. (2019). Visual analytics. In *Data science and visual computing. Advanced information and knowledge processing*. Cham: Springer.

Erickson, A. S. (2021). Joseph G. Gavin, Jr. and MIT's contribution to aerospace in the Apollo era and beyond. *Acta Astronautica, 181*, 167–189.

Faraway, J. J., & Augustin, N. H. (2018). When small data beats big data. *Statistics & Probability Letters, 136*, 142–145.

Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y., & Liu, Q. (2019). Big data analytics and mining for effective visualization and trends forecasting of crime data. *IEEE Access, 7*, 106111–106123.

Feynman, R. P. (1992). *Surely you're joking, Mr. Feynman!*. Vintage; New edition.

Freihaut, P., & Göritz, A. S. (2021). Using the computer mouse for stress measurement: an empirical investigation and critical review. *International Journal of Human Computer Studies, 145*, 102520.

Friendly, M. (2008). A brief history of data visualization. In *Handbook of data visualization* (pp. 15–56). Berlin, Heidelberg: Springer Berlin Heidelberg.

Gao, R. X., Wang, L., Helu, M., & Teti, R. (2020). Big data analytics for smart factories of the future. *CIRP Annals—Manufacturing Technology, 69*, 668–692.

Gupta, S., Kar, A. K., Baabdullah, A., & Al-Khowaiter, W. A. A. (2018). Big data with cognitive computing: A review for the future. *International Journal of Information Management, 42*, 78–89.

Hallikas, J., Immonen, M., & Brax, S. (2021). Digitalizing procurement: the impact of data analytics on supply chain performance. *Supply Chain Management, 26*(5), 629-646.

Han, H., & Li, M. (2018). Big data analytics for high frequency trading volatility estimation. Springer proceedings in business and economics. In M. Tavana & S. Patnaik (Eds.), *Recent developments in data science and business analytics*, chapter 0, (pp. 351–359). Springer.

IDC. (2021). Data creation and replication will grow at a faster rate than installed storage capacity. According to the IDC Global DataSphere and StorageSphere Forecasts. Available at: https://www.idc.com/getdoc.jsp?containerId=prUS47560321

Ishwarappa, A. J. (2015). A brief introduction on big data 5Vs characteristics and Hadoop technology. *Procedia Computer Science, 48*, 319–324.

Jamsa, K. (2012). *Cloud computing: SaaS, PaaS, IaaS, virtualization, business models, mobile, security and more.* Jones & Bartlett Learning.

Jeswal, S.K., & Chakraverty, S. (2019). Recent developments and applications in quantum neural network: A review. *Archives of Computational Methods in Engineering, 26*(4), 793–807.

Jifa, G., & Lingling, Z. (2014). Data, DIKW, big data and data science. *Procedia Computer Science, 31*, 814–821.

Kahil M. S., Bouramoul A., & Derdour M. (2020). Big data and interactive visualization: Overview on challenges, techniques and tools. In: M. Ezziyyani (eds.), *Advanced intelligent systems for sustainable development (AI2SD'2019). AI2SD 2019. Advances in intelligent systems and computing*, vol 1105. Cham: Springer.

Khan, N., Naim, A., Hussain, M. R., Naveed, Q. N., Ahmad, N., & Qamar, S. (2019). The 51 V's of big data: Survey, technologies, characteristics, opportunities, issues and challenges. In *COINS* 19–24. ACM.

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal, 80*, 463–475.

Kumar, R., & Nagpal, B. (2019). Analysis and prediction of crime patterns using big data. *International Journal of Information Technology, 11*(4), 799–805.

Kumar, A., Shankar, R., & Aljohani, N. R. (2020). A big data driven framework for demand-driven forecasting with effects of marketing-mix variables. *Industrial Marketing Management, 90*, 493–507.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, e253.

Lammi, M., & Pantzar, M. (2019). The data economy: How technological change has altered the role of the citizen-consumer. *Technology in Society, 59*, 101157.

Laudon, K. C. (2019). *Management information systems: Managing the digital firm.* Prentice Hall.

Leung, C. K. (2021). Data science for big data applications and services: Data lake management, data analytics and visualization. In W. Lee, C. K. Leung & A. Nasridinov (Eds.), *Big data analyses, services, and smart data. BIGDAS 2018. Advances in intelligent systems and computing*, vol 899. Singapore: Springer.

Levi, M., Allouche, Y., & Kontorovich, A. (2018). Advanced analytics for connected car cybersecurity. In *2018 IEEE 87th vehicular technology conference (VTC spring)* (pp. 1–7). IEEE.

Liesaputra, V., & Witten, I. H. (2012). Realistic electronic books. *International Journal of Human-Computer Studies, 70*, 588–610.

Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing, 115*, 134–142.

Longhi, L., & Nanni, M. (2020). Car telematics big data analytics for insurance and innovative mobility services. *Journal of Ambient Intelligence and Humanized Computing, 11*(10), 3989–3999.

Lucker, J. K. (1994). The evolution of research library buildings. *Computer Methods and Programs in Biomedicine, 44*(3–4), 155–160.

Lyko, K., Nitzschke, M., & Ngomo, A. C. N. (2016). Big data acquisition. In *New horizons for a data-driven economy* (pp. 39–61). Cham: Springer International Publishing.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182.

MIT (1969). Background on the guidance & navigation system for the Apollo Spacecraft. (MIT Instrumentation Laboratory). Available at: https://wehackthemoon.com/sites/default/files/2019-03/Moon%20Show%20%281%29.pdf

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning.* MIT Press. Available at: https://mitpress.mit.edu/books/foundations-machine-learning

Monino, J. L. (2021). Data value, big data analytics, and decision-making. *Journal of Knowledge Economy, 12*, 256–267.

Oak Ridge National Laboratory (2018). New discoveries within 'SIGHT'. Available at: https://www.olcf.ornl.gov/2018/02/07/new-discoveries-within-sight/

Oak Ridge National Laboratory (2019). Laser focus sheds light on two sources of nanoparticle formation. Available at: https://www.olcf.ornl.gov/2019/04/03/laser-focus-sheds-light-on-two-sources-of-nanoparticle-formation/

Özemre, M., & Kabadurmus, O. (2020). A big data analytics based methodology for strategic decision making. *Journal of Enterprise Information Management, 33*(6), 1467–1490.

Peeters, M. M. M., van Diggelen, J., van den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human–AI society. *AI & Society, 36*, 217–238.

Pelteret, J.-P., Walter, B., & Steinmann, P. (2018). Application of metaheuristic algorithms to the identification of nonlinear magneto-viscoelastic constitutive parameters. *Journal of Magnetism and Magnetic Materials,* 464116–464131.

Pesce, D., Neirotti, P., & Paolucci, E. (2019). When culture meets digital platforms: value creation and stakeholders' alignment in big data use. *Current Issues in Tourism, 22*(15), 1883–1903.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science (New York, N.Y.), 331*(6018), 703–735.

Rolls Royce. (2021). What is R2 data labs? Available at: https://www.rolls-royce.com/products-and-services/r2datalabs.aspx

Ross, A., Banerjee, S., & Chowdhury, A. (2020). Security in smart cities: A brief review of digital forensic schemes for biometric data. *Pattern Recognition Letters, 138*, 346–354.

Saggi, M. K., & Jain, S. (2018). A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing & Management, 54*(5), 758–790.

Sartal, A., Carou, D., & Davim, J. P. (2020). *Enabling technologies for the successful deployment of Industry 4.0*. CRC Press.

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question?. *Nature Reviews Neuroscience, 22*, 55–67.

Schmarzo, B. (2013). *Big data: Understanding how data powers big business*. Wiley.

Schroder, A., Lawrence, T., Voets, N., Garcia-Gonzalez, D., Jones, M., Peña, J. M., & Jerusalem, A. (2021). A machine learning enhanced mechanistic simulation framework for functional deficit prediction in TBI. *Frontiers in Bioengineering and Biotechnology, 9*, 50.

Shamout, M. D. (2020). Supply chain data analytics and supply chain agility: A fuzzy sets (fsQCA) approach. *International Journal of Organizational Analysis, 28*(5), 1055–1067.

Shukla, A. K., Yadav, M., Kumar, S., & Muhuri, P. K. (2020). Veracity handling and instance reduction in big data using interval type-2 fuzzy sets. *Engineering Applications of Artificial Intelligence, 88*, 103315.

Sun, X., Yue, L., Yu, L., Shao, H., Peng, X., Zhou, K., Demoly, F., Zhao, R., & Qi, H. J. (2021). Machine learning-evolutionary algorithm enabled design for 4D-printed active composite structures. *Advanced Functional Materials*, 2109805.

Talpes, E., Sarma, D. D., Venkataramanan, G., Bannon, P., McGee, B., Floering, B., Jalote, A., Hsiong, C., Arora, S., Gorti, A., & Sachdev, G. S. (2020). Compute solution for Tesla's full self-driving computer. *IEEE Micro, 40*(2), 25–35.

Termanini, R. (2020). Chapter 5—The digital universe with DNA—the magic of CRISPR. In R. Termanini (Ed.), *Storing digital binary data in cellular DNA* (pp. 97–137). Academic Press.

Thompson, M. P., Wei, Y., Calkin, D. E., O'Conner, C. D., Dunn, C., Anderson, N. M., & Hogland, J. S. (2019). Risk management and analytics in wildfire response. *Current Forestry Reports, 5*(4), 226–239.

van der Aalst, W., Hinz, O., & Weinhardt, C. (2019). Big digital platforms growth, impact, and challenges. *Busisness & Information Systems Engineering, 61*(6), 645–648.

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change, 126*, 3–13.

Wang, Y. C., Zhang, Q., Lin, F., Goh, C. K., & Seah, H. S. (2019). PolarViz: A discriminating visualization and visual analytics tool for high-dimensional data. *The Visual Computer, 35*, 1567–1582.

Wang, H., Li, Y., Khan, S. A., & Luo, Y. (2020). Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. *Artificial Intelligence in Medicine.*

WEF (2021). COVID-19 has accelerated digital transformation - here's how companies can adapt. Available at: https://www.weforum.org/agenda/2021/03/covid-19-accelerated-digital-transformation-how-companies-can-adapt/

Witten, I. H., Frank, E., & Hall, M.A. (2011). *Data mining practical machine learning tools and techniques third edition*. Morgan Kaufmann Publishers.

Xinhua, E, Jing, H., Yasong, W., & Lianru L. (2013). Big data-as-a-service: Definition and architecture. In *15th IEEE international conference on communication technology* (pp. 738–742).

Yang, Z., Yu, C. H., & Buehler, M. J. (2021). Deep learning model to predict complex stress and strain fields in hierarchical composites. *Science Advances, 7*(15).

Zhang, X. D. (2020). Machine learning. In: A matrix algebra approach to artificial intelligence. Springer: Singapore.

Zheng, Y., Sutharshan R., Christopher L., & Palaniswami, M. (2014). Smart car parking: Temporal clustering and anomaly detection in urban car parking. In *2014 IEEE ninth international conference on intelligent sensors, sensor networks and information processing (ISSNIP)* (pp. 1–6). IEEE.

Zhirov, A. O., Zhirov, O. V., & Shepelyansky, D. L. (2010). Two-dimensional ranking of Wikipedia articles. *European Physical Journal B: Condensed Matter and Complex Systems, 77*, 523–531.

Zhong, R. Y., Newman, S. T., Huang, G. Q., & Lan, S. (2016). Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspective. *Computers & Industrial Engineering, 101*, 572–591.

Zhu, Q. B., Li, B., Yang, D. D., Liu, C., Feng, S., Chen, M. L., Sun, Y., Tian, Y. N., Su, X., Wang, X. M., Qiu, S., Li, Q. W., Li, X. M., Zeng, H. B., Cheng, H. M., & Sun, D. M. (2021). A flexible ultrasensitive optoelectronic sensor array for neuromorphic vision systems. *Nature Communications, 12*, 1798.

Modern Machine Learning: Applications and Methods

Xabier Cid Vidal, Lorena Dieste Maroñas, and Álvaro Dosil Suárez

Abstract Machine Learning (ML) is now omnipresent in most fields of human knowledge. Despite this, it remains mysterious and, to some extent, a black box of statistical methods many have heard of, but few know in detail. Although arguably some minimal degree of experience in statistics is needed to master ML, it is not impossible at all to understand many important concepts with no mathematical formalism. In this chapter, we aim at providing a description as complete as possible of what ML is and which is the main methodology used nowadays, from the most basic methods, such as support vector machines, to the very popular neural networks, which appear in some of the trendiest applications of ML. The second part of the chapter will revolve around how ML is used in different areas of knowledge, from industry to basic science. This will be done, once again, with no mathematical formalism, but instead referring to the methods presented in the first part of the chapter. Our goal is that, once a reader has gone through the full chapter, they have a basic idea of what happens behind the scenes when a ML algorithm is run, which are the main types of algorithms that one can use, and they are familiar with the extremely wide range of applications of ML in the modern world.

1 Machine Learning in Modern Days

It does not matter whether you work in industry, in science or as an engineer, Machine learning (ML) is becoming ubiquitous in almost every field associated with human progress (Brink, 2016). ML has been among us from many years ago, in applications such as image recognition, product recommendation, fraud detection... even

X. Cid Vidal (✉) · L. Dieste Maroñas
Instituto Galego de Física de Altas Enerxías (IGFAE), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain
e-mail: xabier.cid@usc.gal

L. Dieste Maroñas · Á. Dosil Suárez
Triple Alpha Innovation, Cuns, 4, 15286 Outes, A Coruña, Spain

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8_2

without us knowing it, and now is one of the boosting technologies of the 4th industrial revolution. The main goal of this chapter is making you understand why this is the case and quickly guiding you through the main methodology of ML, while providing examples of applications in several areas of knowledge. ML is defined by the Cambridge dictionary as “the process of computers changing the way they carry out tasks by learning from new data, without a human being needing to give instructions in the form of a program”. While this is a relatively simple definition, as we shall see it gathers a lot of important information about how ML works. For the moment, let us simply say ML is a group of powerful statistical techniques which, using the simplistic concept of “learning through examples”, can achieve tasks as amazing and diverse as allowing a car to drive itself, optimizing the performance of an assembly line or allowing the discovery of a new subatomic particle.

You might have used ML in the past, even without being aware of it. Imagine you have a doughnut production line, and you want to eliminate the faulty ones. You look at a few examples of those you consider not suitable for selling, and you visually find some patterns: they have a blackish color, their hole is not as perfect as it should be, or their size is not appropriate. Based on these criteria, you write down some guidelines so your workers can discriminate against those doughnuts you do not like to sell. As simple as it is, this is ML. You have learned from examples on your own data, found patterns, and developed a method to select what you want in independent data sets! This kind of task is called “classification”, and it will be reviewed in the next section. As a different approach, say you want to study how the average temperature in different world regions of the same longitude from the northern hemisphere changes with latitude at the same time of the year. As a rule of thumb, you could expect the temperature goes down with latitude (the more to the North, the lower the temperature). A scatter plot of different measurements (temperature vs latitude) can be very helpful and easily show such a correlation. You take one step further, assume the correlation is linear, and perform a regression analysis to model it. Using this model (no need to enter in this example on how accurate this is) and given a latitude or temperature, you can estimate which is the other expected magnitude in the pair. And surprisingly, or not again, that is ML! You have taken some data, used some simple statistics to understand it, and provided a model you can use to produce independent predictions. As we shall see, this is what we generally call “regression”.

This chapter is divided into two main parts. In the first one, we will provide a definition of the types of learning used in ML, such as supervised and unsupervised learning. Furthermore, we will briefly summarize some of the most popular methodologies, without providing any mathematical detail. This will include concepts such as boosting, support vector machines, or the very popular artificial neural networks (and associated methods). In the second part, we will cover the main uses of ML in many different aspects of the modern world, from engineering, manufacturing, and finance to user interface, medicine, and science.

2 How Can We Teach a Machine to Learn?

When we talk about ML we are referring to the process through which a computer learns how to solve a problem. This learning can be classified in two main different types of learning: supervised and unsupervised (Géron, 2019).

2.1 Supervised Learning

In supervised learning a human directs the algorithm learning by giving it examples of the problem and the desired solutions (Pedregosa et al., 2011). Basically, the person must provide to the algorithm a dataset well labeled, i.e., the training examples must be tagged with the right answer, so the algorithm can learn what are the properties that define each label are. The right answer can be a class from a classification problem or a specific number as a result of a regression. After that, the supervised learning algorithm examines the training data so, when the machine is provided with a new set of unlabeled data, the algorithm will be able to produce a correct outcome.

There are dozens of supervised learning algorithms, but mostly all of them can be included in one of these groups: classification or regression. The goal of classification is to label an input, although regression aims to predict a quantity.

2.1.1 Classification

Classify means essentially to group some elements by their properties. Humans like to classify everything since it creates order that helps us to understand the world. So, this is a very interesting feature of ML for us. In ML classification problems, the goal will always be to group the inputs in different categories depending on their characteristics. A typical example of classification is the email spam detection. In this problem, there are only two different output classes, spam or legitimate. Since we only have two options, this is called a binary classification problem. Every time we receive a new mail in our inbox, the ML algorithm must classify it as spam or not-spam. However, we can find classification problems with any number of classes. For example, if we want to classify a book by its genre, we have a lot of different output classes: romance, thriller, biography, or adventure.

There are tens of classification algorithms to choose from, but here we are going to mention only some of the most important ones: Logistic Regression, Decision Trees or Neural Networks (NN), and Support Vector Machines (SVM).

2.1.2 Regression

Unlike classification, in regression problems the result is a number. In this kind of problem, the goal is to predict a numerical property of the input based on its properties. One of the most used examples to explain this method is the house price prediction. As its name indicates, the algorithm must predict the market price of a house based on some of its inputs such as its size, date of construction, location, and many others. Some of the main methods for regression are lineal and no-lineal regression, Decision Trees Regression or Neural Networks, and Support Vector Machines.

2.2 Unsupervised Learning

Opposite to supervised learning where a human must teach the algorithm how to perform, the most defining characteristic of an unsupervised learning algorithm is that it must figure the problem and find the solution by itself. That is, it must recognize the data structure and what the outcome of the algorithm is supposed to be. These kinds of algorithms are very useful to discover hidden patterns in data or to feature learning and, also, since they do not need human intervention, they can improve their results over time by themselves.

One of the main examples of unsupervised learning is clustering. The main goal of a clustering model is to group the elements by its features. Objects in the same groups will be similar, while objects in different groups are dissimilar. Two elements of the same group do not need to have identical values of a feature but will be like a third element of another group.

The ML models create a spatial representation of the features and then evaluate the similarity of the elements by the distance between them. The smaller the distance, the similar the objects are and, hence, they probably belong to the same group. In the end, the result is that the distances between neighbors within one cluster are smaller than between objects from different clusters.

2.3 Others

2.3.1 Reinforcement Learning

We also must mention a fourth type of learning called reinforcement learning (RL) (Osiński & Budek, 2018). In this case, the algorithm must operate into an environment without any instructions of how to do it. Supervised and unsupervised algorithms need to be fed with data first; nevertheless, this step is completely skipped in reinforcement learning. Instead, the data is generated from trial and error during training, and it is tagged at the same time.

The goal of these algorithms is to perform a task as well as they can. The algorithm is connected to an environment where it can accomplish the task, and the only interaction with the trainer is that the algorithm receives a feedback or reward at certain times every time it gives a try. So, the aim of the algorithm is to increase the reward as much as it can, creating long-term strategies.

There are two kinds of reinforcement learning methods: positive and negative. The positive reinforcement maximizes the actions that increase the reward. This type of reinforcement maximizes performance and sustains change over a longer period, but excessive reinforcement may lead to over-optimization, affecting outcomes. Negative reinforcement, on the other hand, is defined as any alteration in behavior caused by a negative condition that should have been avoided. Thus, the disadvantage of this method is that it provides enough to achieve the minimum objectives, but does not provide the optimal solution. Some applications of RL can be:

- Robotics for industrial automation.
- Creation of training systems that provide custom instructions.
- Business strategy planning.
- Aircraft control and robot motion control.
- Machine Learning and data processing.

3 Different Algorithms and Methods Used in ML

At the moment of writing this book, we can count tens of ML methods, being even difficult to enumerate all of them. To avoid leaving the scope of the book, we are going to review the main methods that are being used nowadays.

Before we get down to business, some terms are worth describing for a better understanding of the chapter. A dataset is a collection of objects, and an object can be anything we want to use to solve a problem. For example, different flowers, emails, different animals… The objects are described by features. In the flower example, some features would be the color, the size, and different properties we can use to describe them. These features can be of different types, and the most used ones are numerical, categorical, and binary. The numerical type is used when the features are described by numbers, as, for example, the age of a flower. Categorical features usually describe a property that is not a number, as the color of the flower. Finally, we use a binary feature when we want to describe a property with only two options, like true or false.

All the objects and their features together form what we call a dataset. During the ML model training, the dataset will be divided into two smaller sets, the train set, those objects we will use for training our model, and the test set, those objects we will use to "test" the final model.

Once the two datasets with their objects and features are defined, the next step is to describe the problem we want to solve and, afterwards, the algorithm we will use. The algorithm is a mathematical application that will map our dataset to the solution we want to achieve. There are a lot of different algorithm options we can choose from, some of the most known ones are Logistic Regression, Support Vector Machines, Neural Networks, and many others.

Once we run the algorithm on data, we create the ML model, i.e., the model is the algorithm parametrized after the algorithm training. It represents the rules, numbers, and any other data structures required to make predictions. The model gives us an approximation of the result, the closer the result is to the real one is a measurement of how good our model is. One of the most difficult parts of creating a ML solution is choosing the right algorithm, which means applying the right assumptions about the data we have.

After achieving a good enough result during training, it is time to evaluate the model and check how good it really is. In order to do that, we need to evaluate it using a set of objects different from those used for training. This dataset is called the test set.

At this moment, some training errors can arise. Two of the most common problems are overfitting and underfitting. Overfitting means that the model can approximate the training data almost perfectly, although when tried with new data the accuracy decreases. In this case, the model is said to have high variance. The reason is that the model learns the noise in the training set. This noise can obscure the true relationship between features and the response variable. Overfitting is more likely when there are many features available, or a complex model is used. The more the features, the bigger the chance of discovering a fake relationship between the features and the result. Complex models develop more complex hypotheses about the relationship between features and the result. When a model underfits it is not able to fit the training data and it is said to have high bias. This means that the model is not complex enough, in terms of the features or the type of model being used.

3.1 Logistic Regression

Logistic regression is an ML technique from the field of statistics. From its name one might think that it is an algorithm to be applied in regression problems, however it is a method for classification problems, in which a binary value between 0 and 1 is obtained. Thus, logistic regression allows to establish the possible relationship between a dependent variable with one or several independent variables through a logistic function that determines the probability that the dependent variable is related to the independent variables (Osiński & Budek, 2018). The independent variables represent the features of the objects we want to classify, and the dependent the variable, the class the object belongs in.

The linear regression model has a quantitative variable as the output variable, nevertheless for classification we need a qualitative output. To exemplify this, let us think of a variable called "severity" that indicates if a patient's condition in a hospital is "serious" or "not-serious". Thus, we have two groups, 0 = "not-serious" and 1 = "serious". Logistic Regression manages to transform the output variable with the

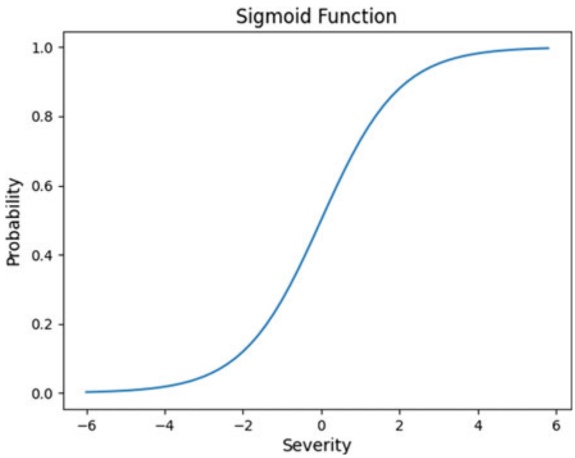


Fig. 1 Graphic representing a sigmoid function

logistics operator, also called sigmoid function. This mathematical operator converts the independent feature in a probability, ranging between 0 and 1 and representing how likely an instance is of being 0 = “not-serious”. The sigmoid function is an S-shaped curve with values between 0 and 1 (see Fig. 1).

This probability must be translated into binary values, for which a threshold value is used. For probability values above the threshold value the statement is true and below it is false. A true positive is an object that was classified in the correct class ‘0’ and a false positive is an object that was classified in the class ‘0’ but it belonged in the class ‘1’. With false negatives the idea is the same, an object that belongs to the class ‘0’ is wrongly classified as class ‘1’. In this way we can use the same structure as linear regression. We are simply converting the response variable, which is qualitative, into a probability, which is quantitative.

Logistic regression is a technique widely used by data scientists because of its efficiency and interpretability. In addition, it does not require extensive computational resources for training or execution. The performance of logistic regression, like linear regression, is better when using attributes related to the output. It is also important to eliminate features that show great multicollinearity with each other. Therefore, the selection of these features before the training of the model is a key. Because the expression that makes the decision is linear, the model is not able to solve non-linear problems directly and it is better to use other models such as decision trees.

3.2 Support Vector Machines

Support Vector Machine (VVM) can be defined as a supervised method mostly used to solve classification problems, but also applied in regression problems. In this last case, we talk about Support Vector Regression (SVR). When working on classification problems the algorithm learns how to separate classes creating decision

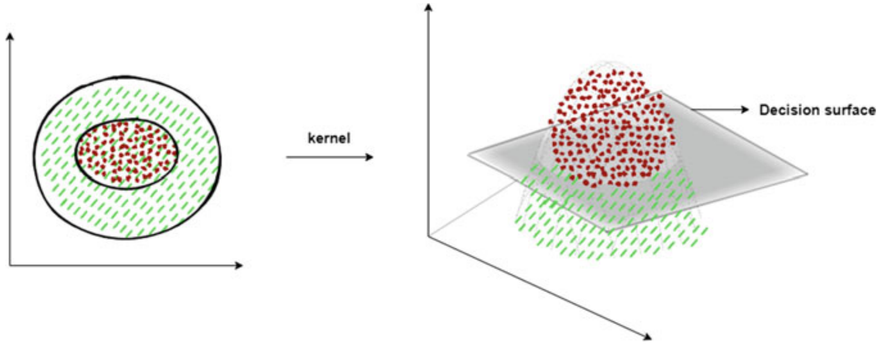


Fig. 2 Example of a kernel transformation. (Left) the data as we have it in a two dimensions space. (Right) The data after some transformation to be linearly separable in a three-dimension space

boundaries or hyperplanes (Scikit-learn, n.d.). The points that define the maximum margin of separation from the hyperplane are called support vectors. They are called vectors, instead of points, because they have as many elements as there are dimensions in our input space.

Sometimes it is impossible to find a hyperplane to separate two classes, so these two classes are not linearly separable. To tackle this problem, we have to use a kernel. This consists of inventing a new dimension in which we can find a hyperplane that separates the classes. In Fig. 2 you can see an example of how the kernel works.

Basically, we transform the two-dimensional space without a linear separation, into a three-dimensional space. In this new space we clearly observe a plane separating the two classes.

As the reader may record, we said that a regression problem is based on looking for the curve that models the trend of the data and, according to it, predicting any other data in the future. In principle this definition is not very compatible with the support vector machines but doing some simple changes we can get it ready. The SVR uses the same principle of SVM and if the problem is not linear, it adds a new dimension to try to find a linear separation.

3.3 Decision Trees and Random Forests

A decision tree is a method to separate a set of objects into several distinct subsets through binary decisions about the properties of its elements. Each binary decision consists of a comparison involving one or more variables, and it is taken in the Decision nodes (Gupta, 2017). The output of the node are two new decision nodes or two terminal nodes and depending on the result the child goes to the left or to the right. The classification starts at the root of the tree, in the node called Root node and finishes on the Terminal nodes that do not split and that determine the result of the operation.

The main advantage of trees is that they represent "rules" which can be understood by humans with the advantage that the knowledge is generated by the tree itself and not based on the premise of an expert on the subject. Another advantage of this algorithm is that it is easy to use, everybody can understand the way it works and plan a model to solve problems. It is versatile, since a lot of different problems of different fields can be solved using this modeling. Moreover, compared with other algorithms it requires less effort for processing the data, this means no normalization, not filling missing data and not scaling among others. However, this type of solution also has disadvantages, mainly because it requires long training times for the models and is often unsuitable for predicting continuous values, and is mainly used for classification problems.

There are a lot of options in order to improve the algorithm and achieve a better result for the model, one of the most known ways of improving a Decision tree is the combination of several of them called ensemble learning and resulting in a new method called Random Forest (Glen, 2019). Ensemble-type methods are made up of a group of predictive models that allow better precision and model stability to be achieved since the group compensates for the errors in the predictions of the individual ones. Also, Random Forests use a technique called bagging, which means that different trees see different portions of the data. This way we can prove that each tree is different from the others. At the end we have many versions of the algorithm trained on different subsets of the dataset so that the biases cancel each other. Although each decision tree has a high variance, when combining them the total variance is low as each decision tree is perfectly trained and therefore the outcome does not depend on one decision tree but on multiple decision trees.

An example of the process can be seen in Fig. 3, where a schema of how a Random Forest algorithm with three Decision Trees would work is shown.

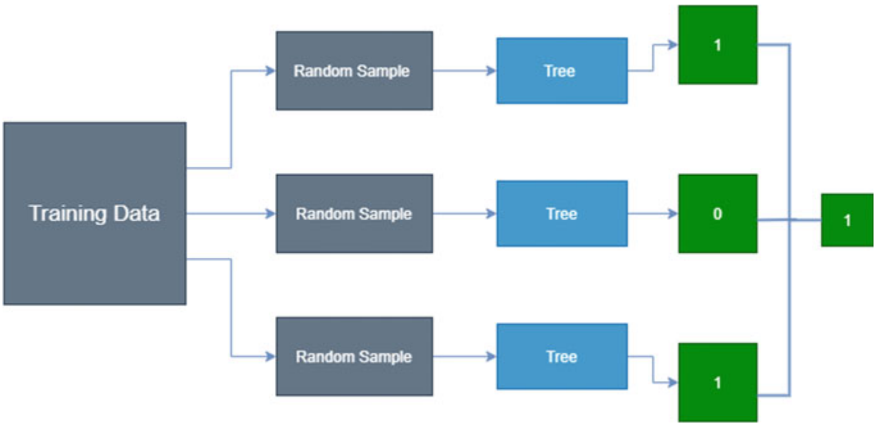


Fig. 3 Schema of how Random Forests works. It combines three different regular Decision Trees to obtain the final result using soft-voting

Random Forest can be used both for classification and regression solutions. The process is similar for both, being the combination of the result the main difference. In classification problems, the results of decision trees are often combined using *soft-voting*; however, we can also come up with our own way of combining the results of a random forest. In soft-voting, more importance is given to some of the decision trees of the forest. Regarding regression problems, the most common way to combine the results of the decision trees is by taking their arithmetic mean.

3.4 Boosted Decision Trees

Another improvement of the performance of a Decision tree can be what we call BDT or Boosted Decision Tree. This is another ML technique used for regression analysis and for classification problems. The name boosting is an approach of creating a highly accurate capability of prediction by combining weak and imprecise predictors. BDT algorithms consist of a set of decision trees, each with an assigned weight. Each tree is created iteratively using the output of the previous tree, and the tree's output is given a weight relative to its accuracy. To calculate the final output, the outputs of the trees are linearly combined with their weights. After each iteration every data sample is weighted proportionally to the frequency of its misclassification. The final goal is to minimize the loss function. A schema of how the model uses the residuals of the previous model, and the combined solution is shown in Fig. 4.

This new method brings some new parameters that we will have to take into consideration in order to create the best possible model. The most important ones are:

- Loss function: the loss function indicates the difference between the real data and the prediction.
- Learning rate: the learning rate indicates the step to adjust data weights in each iteration. The smaller the more precise, but also slower. Normally this parameter takes values close to 0.1.
- Subsample size: as data samples are randomly selected in each iteration, the subsample size is the parameter that indicates how many samples to train in each new tree.
- Number of trees: the total number of trees that can be created to solve the problem. Usually a big number is better, but it could lead to overfitting.

As every other algorithm BDT has advantages and problems. We can highlight that this algorithm is fast, for both training and prediction, easy to tune and it is not sensitive to scale, allowing a mix of numerical or categorical features. Also, it has good performance, and it is very commonly used so there is a strong community behind it. On the other hand, BDT is an algorithm sensitive to overfitting and noise.

So far, Random Forest and BDT seem similar, but they have two important differences. The first one is related to how the trees are built. While random forests build each tree independently, the BDTs build one tree at a time, and it introduces weak

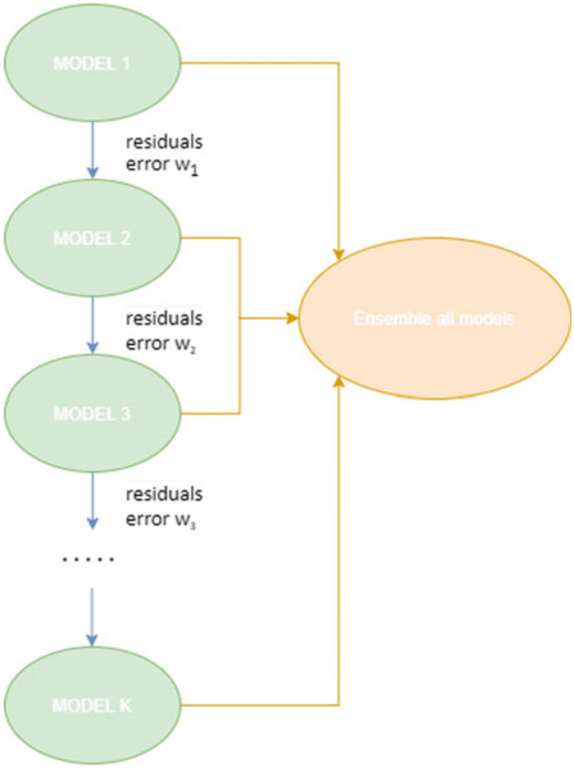


Fig. 4 Schema of how boost decision tree works. Example using k models combined

learners to improve the deficiencies of the previous ones. The second one is the way they combine the results. Random forests combine the results at the end of the process, while BDT combines results after each iteration. Hence, BDT can result in higher performance. Nevertheless, the final performance will depend on the problem. In a problem with very noisy data, the BDT may not be a good option as it can result in overfitting. Also, it is usually more difficult to tune than Random Forests.

There are many types of boosting algorithms, some of the more common ones are AdaBoost or Adaptive Boosting, Gradient Boosting, and XGBoost. The main difference among the algorithms is the loss function they use to calculate the weights of the trees and the data.

3.5 Neural Networks

Neural Networks try to mimic the network of neurons in a human brain and its behavior. Basically, the structure of a neural network is a bunch of neurons connected among them and working with a common goal, without having individually a specific task. The neurons can be grouped in layers that are connected among them. Typically,

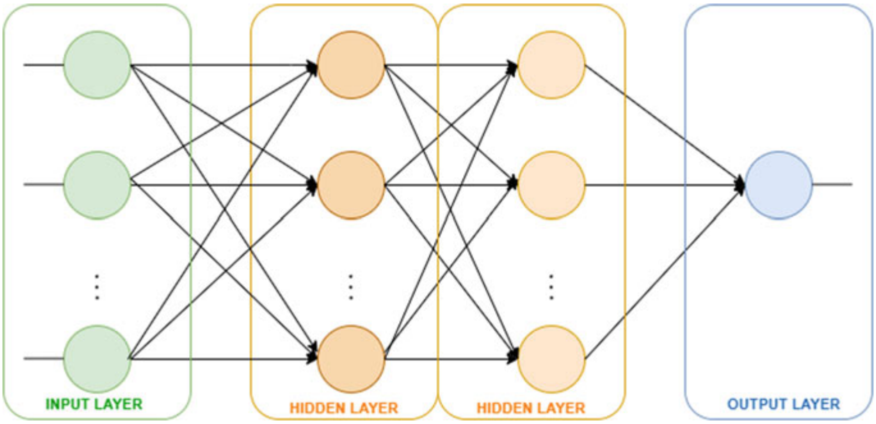


Fig. 5 Neural Networks graph representation. Example with 2 hidden layers

the NNs are formed by three layers. An input layer, with neurons representing the input fields (Goodfellow et al., 2016). One or more hidden layers, and an output layer, with one or more neurons representing the result. An important characteristic of neural networks is that every input to a neuron is weighted, which is a crucial aspect when training the network. An example of a diagram of a neural network can be seen in Fig. 5.

3.5.1 Deep Neural Networks

One of the multiple possible classifications of the NN is by the number of their hidden layers. Using this classification, we find Shallow neural networks that have only one hidden layer between the input and output, and Deep Neural Networks (DNN), that have multiple hidden layers. One example of DNNs is the Google LeNet model for image recognition that has 22 layers (Alake, 2020).

In DNNs, each network layer can learn increasingly complex features of the data. Each layer integrates a deeper level of knowledge. A neural network with five layers can learn more complex features than one with only two layers. In this kind of networks, the hidden layer is divided into two phases. The first is called dense layer, and it applies a nonlinear transformation to the input data. The second phase improves the model with a derivative function, in what is called the activation layer. The neural network repeats these two steps many times until the result is similar enough to the one desired. Each repetition of this two-phase is called an iteration. An example diagram of a neural network can be seen in Fig. 6.

Deep neural networks provide higher accuracy in many tasks compared to any other method, for example, in object detection or speech recognition. Also, they can learn autonomously, without any human knowledge transference.

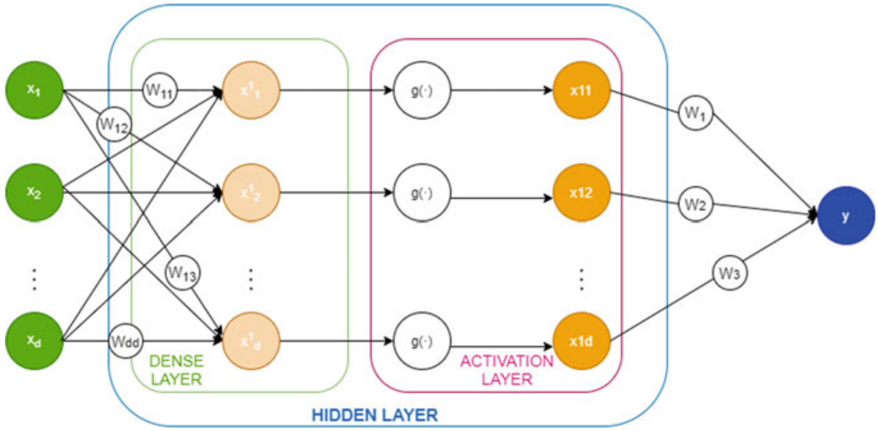


Fig. 6 Neural Networks graph representation of the hidden layer. The Wij are weights that represent the importance of every input when feeding the next neuron. Example with d features

Deep learning is called this way because it makes use of deep neural networks. This type of learning can be supervised, semi-supervised, or unsupervised.

3.5.2 Convolutional Neural Networks

A Convolutional Neural Network (ConvNet/CNN) was born as a necessity of improving the DNN when they interact with images (IBM Cloud education, 2020).

A color image can be represented as a mix of three matrices, one for each main color, Red, Green, and Blue, in which the number of pixels indicate the size of the matrix. For example, a color image with 34×34 pixels is represented as three matrices of 34×34 units each. In Fig. 7 we can see how a regular picture is represented by three images, Red, Green, and Blue.

This means that images are high-dimensional objects that demand powerful techniques for efficient processing. With all this, CNN is a type of Artificial Neural Network that using supervised learning imitates the visual cortex of the human eye. In CNNs hidden layers are specialized in one specific task and are sorted hierarchically. The first layers can recognize simple shapes such as lines or curves. Deeper layers can detect much more complex shapes such as faces or landscapes.

CNN requires much lower pre-processing compared to other algorithms. This is what we call the “distinctive processing” of a CNN. That is, the so-called “convolutions”. These consist of multiply scalar, a group of nearby pixels from the input image by a kernel. The kernel goes through all the input neurons creating a new hidden layer. In Fig. 8 a schema of how an image is treated by a CNN is shown.

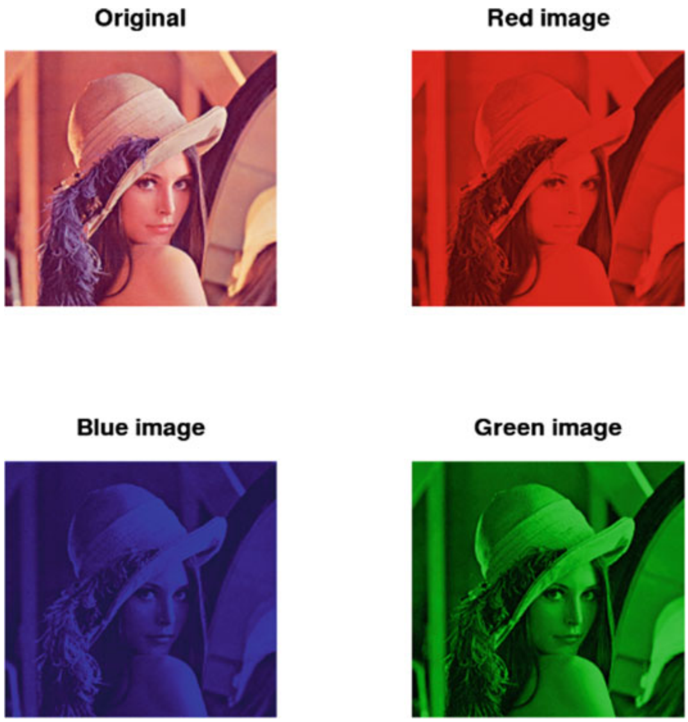


Fig. 7 Red, Green, and Blue example of how an image is shown in the RGB representation

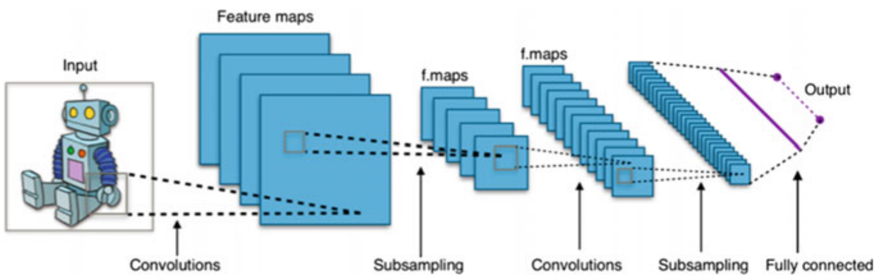


Fig. 8 CNN schema of how a convolutional neural network works. This file is licensed under the Creative Commons Attribution-Share Alike 4.0 International license

3.5.3 Recurrent Neural Networks

So far, we have seen networks whose activation function only acts in forward direction, from the input layer to the output layer, that is, they do not remember previous values. A Recurrent Neural Network (RNN) is similar, but includes connections that point backwards, a sort of feedback between neurons within the layers. These networks are called recurring because they perform the same task for each element

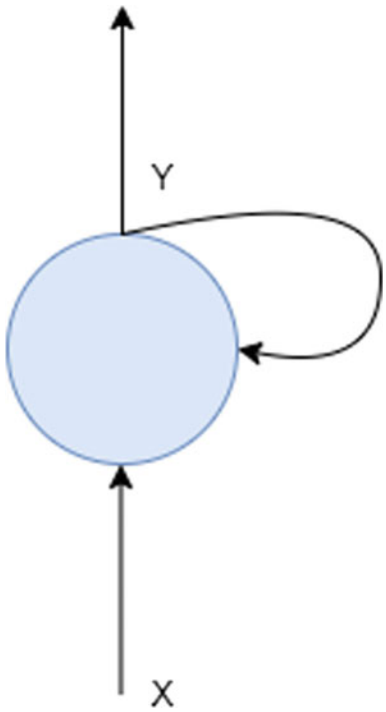


Fig. 9 Single RNN neuron with feedback of the output

of a sequence, and the output depends on the calculation explained above. To see how this works let us imagine the simplest possible RNN. This is the case of a single neuron that receives an input, generates an output, and sends back that output to itself. This idea is shown in Fig. 9, in which a single neuron with a regular input and the previous output are shown as the inputs for the process.

At each instant (also called timestep in this context), this recurring neuron receives the x input from the previous layer, as well as its own output from the previous time instant to generate its output y. Following this same idea, a layer of recurring neurons can be implemented in such a way that, at each instant of time, each neuron receives two inputs, the corresponding input from the previous layer and in the output from the previous instant of the same layer.

Given that the output of a recurrent neuron at a certain moment is a function of the inputs of the last moments, we could say that these kinds of neurons have memory. The cell state is preserved over time in an internal memory called memory cell. This internal memory makes the RNNs a very interesting method to apply on ML problems involving sequential data. This internal memory allows RNNs to remember relevant past input information, which allows them to make better predictions and keep contextual information. These algorithms are used for temporal problems, such as natural language processing, speech recognition, and language translation. Examples of applications are voice search, Alexa and Google Translate, detect fraudulent credit-card transactions, etc.

One of the disadvantages of a simple RNN is that it has a short memory. However, it has shown remarkable success in natural language processing, especially with its Long Short-Term Memory (LSTM) variant that can look back longer than RNN and, therefore, solves the problem of the short memory.

3.5.4 Natural Language Processing

Natural Language Processing (NLP) is an area that provides machines with the ability to comprehend the human language. NLP makes it possible for computers to read a text, listen to a spoken voice, extract the meaning, determine which parts are important, and even measure the sentiment. Virtual assistants or chatbots are one of the best-known utilities of the NLP, but they are not the only ones. In addition, it is important to understand that NLP does not give a chatbot intelligence, it only gives it the ability to process and generate human language. To provide intelligence to a virtual assistant, the use of systems such as neural networks is necessary.

NLP has many parts and many steps to succeed and being able to understand the human language. Usually, NLP starts dividing the text into elements (phrases, words, etc.) and trying to understand the relationships between them. One of the most popular models to divide the text in words is called Bag of Words. This model is very used. It counts all words in a text and creates an occurrence matrix neglecting word order and grammar. In this process it is also useful what is called Tokenization. This consists of cutting a text into pieces called tokens and removing some characters without interest for the analysis, as for example, punctuation characters. This part is what we call Natural language understanding (CLN or NLU). It is the part of natural language processing that is responsible for interpreting a message and understanding its meaning and intention, just as a person would. In order to get the system working, you need datasets in the specific language, grammar rules, semantic theory and pragmatics, etc. In Fig. 10 we can see one of the simplest examples of tokenization.

Another popular part of NLP is Speech to text or STT. It is based on the conversion of audio to text, and it is a task to value the audios, which once converted into texts can be processed with other NLP techniques. Once processed it is possible to return an audio using the text to audio conversion (Text to Speech or TTS). Both tasks, STT and TTS, have become truly relevant with conversational systems with a prominent level of quality.

To sum up, the applications of NLP are huge, and the fields of application increase every day. Let us mention some examples:

```
Can you read this text?
['Can', 'you', 'read', 'this', 'text', '?']
```

Fig. 10 Tokenization of the sentence "Can you read this text?"

- Organizations can extract valuable information about customer choices and, mainly, their decision drivers. Also, they can determine the feeling about a product or service by analyzing the sentiment analysis, for example, social media.
- From the email text analysis using NLP, your email provider can classify your emails and stop spam before entering your inbox.
- NLP is also being used in talent recruitment. It permits to identify the applicant skills in the selection phase.
- Automatic Machine Translation is a field of research within computational linguistics that studies systems capable of translating messages between different languages or languages. For example, Google is one of the companies that has invested the most in machine translation systems, with its translator using its own statistical engine.
- Autocorrect and autocomplete text systems also use Natural Language Processing.

More examples of the different uses of NLP are given in the next section of ML applications.

3.5.5 Generative Adversarial Networks

As its name may suggest, generative adversarial network (GAN) algorithm consists of two neural networks competing one against the other. One of the networks, called the generative, produces samples of what we try to create, and the other one, the discriminating network, examines the samples and determines if it fits the requirements. If the samples do not fit these requirements, they are discarded, and the generative is notified about that, forcing it to try again. This process can be repeated thousands or even millions of times until the discriminating network agrees with the result. In this process, the generating network learns what the discriminating network is looking for.

One of the latest and most surprising applications of this technology is the fake human face generators. An example can be the DCGAN from Nvidia. This tool allows to generate hyper-realistic faces that do not correspond to any real person. Another use of this technology is to generate samples of photorealistic images of industrial design, interiors, clothing and accessories, or elements for scenes from computer games.

3.5.6 Autoencoders

The most representative characteristic of autoencoders is that in this kind of networks, the input is the same as the output. Their objective is the generation of new data. Autoencoders compress, or encodes, the input into a lower dimensional latent representation code and then reconstruct the output from it (Dertat, 2017). This type of network consists of three parts, Encoder, Code, and Decoder. The encoder is the input part of the network, and it compresses the input into a space of latent variables.

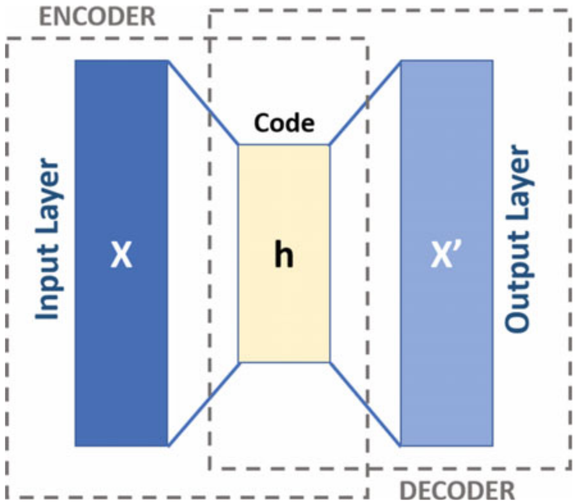


Fig. 11 Autoencoder representation. It has three parts, the encoder, the code, and the decoder. This file is licensed under the Creative Commons Attribution-Share Alike 4.0 International license

The product of this process is the Code. Decoder is the part that tries to reconstruct the input based on previously collected information, the Code. These three parts are connected as it is shown in Fig. 11.

If we observe the code, we will see that at this point the Autoencoder has a compact representation of the input. That is, the data obtained is a compressed version of the input and therefore contains a smaller amount of data. The representation obtained in the code is known as latent space and is the result of training, where the network learns how to extract the most relevant information from the input data. What we hope is that, when we train an autoencoder and copy the input, the code can take on useful characteristics for us. To find this compact representation, the Autoencoder is trained in an equivalent way to a Neural Network. However, in this case the error function used to update the autoencoder coefficients is simply the result of comparing point by point the reconstructed data with the original data. Taking this into account, we can say that the Autoencoder is an example of unsupervised learning, because during the training we do not really define the category to which each entry belongs.

Some of the most popular uses of Autoencoders are dimensionality reduction, removing noise in images or detecting anomalies in a series of data.

3.6 K-Nearest Neighbors

K-Nearest Neighbors (kNN) is the main example of a clustering algorithm. As we mentioned above, clustering is a type of unsupervised ML. This means that in the dataset we just have objects and features, and we do not know the classification of the objects.

These algorithms take a new element and calculate the distance from it to every other element. Depending on the distance to k neighbors the new element is classified into a group or another. The selected group will therefore be the one with the highest frequency with the shortest distances. To decide the group the point belongs to, there are two values that we must define beforehand. The value of k, which is the number of elements per class, and the formula to calculate the distance between the points. The most popular ways to "measure the closeness" between points are the Euclidean distance or the Cosine Similarity (it measures the angle of the vectors, the smaller, and the similar). The value of k is especially important, as this will almost end up defining how many groups are and to which group the points will belong. The best choice for k depends on the data. Normally higher values of k reduce the way noise effects' classification. However higher values of k make the boundaries between classes less differentiable.

Despite its simplicity, it is used in solving a multitude of problems, such as recommender systems, semantic search, and anomaly detection. The main advantage of this algorithm is that it is simple to learn and implement. Its disadvantages are that it is very memory and processing resources (CPU) demanding, since it uses the entire dataset to train each new point. So, because of this, KNN is best suitable for small datasets and with a low number of features.

4 Machine Learning Applications in the Real World

As we already said at the beginning of this book, ML is being used in many and many applications, so it is impossible to condense in a chapter a full comprehensive review of the ML applications. Instead, this chapter aims at showing the palette of current applications, ranging from engineering to fundamental science. We provide significant examples in each area, focusing on the methodology, and relating every topic to the techniques discussed above.

4.1 How Machine Learning Improves Engineering

The work of engineers has not escaped the current revolution, and the way their jobs are done has had to dramatically align with the most modern trends in ML. The introduction of ML and all its tools bring significant improvements to what we are capable of on our own (Smith, 2020). One of the most distinct aspects of ML incorporation in engineering has to do with the workflow management. ML goes well beyond the designing, into data management and interoperability. As one could expect, ML helps to manage engineering data in a more optimal way, effectively adapting in Industry 4.0, solving the problem of big data, and making other advancements easier to deal with (Hulten, 2019). Moreover, ML is also allowing a better communication between different departments to perform different tasks. If computers and systems are capable of learning over time, then they can improve and automatically carry out many daily efforts. In summary, ML is perfect to improve our capabilities in engineering, given as we have seen, it is able to take advantage of computers by imitating our own learning.

We all know engineering problems which solve traditional computational techniques to the limit. Many times, they are solved because there are experts capable of reaching the most appropriate solution, and then checking it with conventional calculation methods. In this context, ML is trying to capture the essence of human cognition to accelerate the resolution of these complex problems. ML has been used intensively in engineering research in recent decades. Newer methods such as pattern recognition, ML, and deep learning are emerging methods in this field of engineering. These emerging techniques could learn complicated dependencies between parameters and variables and hence allow facing more efficiently a variety of obstacles that otherwise could not be dealt with by means of traditional methodology.

One of the main aspects of ML is improving our lives and engineering in general is when it is applied in Big Data (Hasan et al., 2014). For some years the amount of data existent on the Internet has grown exponentially. This means we can find huge amounts of data with different structures, which appear in business at almost any time. The key with this data is not so much the amount of it we have, but rather how companies use it. In this regard, it becomes essential to analyze it in detail to later take good business decisions that, e.g., improve their income. And it is in this part of the process in which ML starts to be useful for engineers. Some examples of uses are:

- Tourism: the tourism industry is essentially about keeping customers happy. However, quantifying this happiness is often hard. Furthermore, the detection of dissatisfaction must be done as quickly as possible, otherwise there might be no time of reverse a bad customer experience. Big data provides handles to companies to tackle these problems, by providing advanced analytics of customer data, helping to prevent potential problems well in advance. Recommendation systems built by NN are one of the main tools Tourism uses to promote and reach the targeted population, to increase the benefits.
- Administration: administration must tackle the challenge of maintaining excellence and productivity with a cost as low as possible. This is even more true for justice.
- Autonomous car: the era of the autonomous car will be a reality in the not-so-distant future. In fact, there are already concrete steps in this direction. Driving without having to do anything at all will become a very common habit in a few years. Facing that horizon, more and more brands announce new research. These are focused on security systems that include the most amazing technologies for when the autonomous car arrives. In this scenario, ML already shines and will be one of the great protagonists. Specially DNN and CNN are used to work with the information the diverse sensors of the car recollect (Amezquita-Sanchez et al., 2020). This information changes from temperature, speed to images of the road, making the use of complex algorithms a necessity.

However, ML not only improves data management and the way Big Data offers solutions. ML can also be applied to electrical and computer systems. ML makes it possible for logic control to create rules that enable machines to react against inputs of very different kind, rather than simple binaries (Aguilar Vidal et al., 2020). In general, it makes it possible for systems to discover different arrangements, find inferences, and understand how to perform tasks with no specific guidance. In addition, ML has a high potential in the field of computational mechanics, given how these methods can solve complex problems through the so-called Internet of Things. This introduces one of the main developments of the area, the time for smart infrastructure, cities, homes, or structures. Smart Cities are not a utopia. In fact, smart cities are a reality whose basis is the use of infrastructure, innovation, and technology to reduce energy consumption, reduce CO_2 emissions, and increase the quality of life of the population. The parameters that serve to qualify a city as Smart City are its commitment to the environment, its urban planning, its public management, its mobility and transport conditions, its efforts to facilitate social cohesion, and human and economic investments to improve its operation. Following the concept of Smart City, the idea of Smart Home is introduced as those that incorporate a system that allows automating many tasks, as well as having full and live control over what happens in our home (Ricquebourg et al., 2007). The principal goal is to incorporate automation to manage air conditioning, lighting, security, audio or video. All these systems are interconnected and can be controlled remotely. Smart homes are also committed to efficient and intelligent consumption, always adjusted to the needs of users.

Finally, we must take the job evolution into consideration. We have a wealth of examples in the past of how innovation usually becomes a new sector of research and work, and this time it will not be different. In this regard, most of our innovations in engineering currently have ML as a key feature. Our workflows will evolve, new fields of study will appear, and new types of industry are to be developed so that existing engineers will need to adapt to. We can already see how data and ML engineers are the new fashion profession in the business and research sector.

However, we still encounter limitations in the use of these emerging methods. These limitations include how difficult it is to select the most optimal ML method, the fact that the repercussions of noisy or incomplete data as well as the efficiency of computation are not considered, or the non-reporting of the accuracy of the data. Other examples are the classification without examining different solutions to increase the yield, and the inadequacy of the presentation of the process to select the best parameters for the ML problem under scrutiny. Yet despite these limitations, ML, pattern recognition, and DL are posited as pioneering methods for increasing the efficiency of many applications in modern engineering, as well as for developing new possibilities.

The automatization achieved by ML models, such as NNs, usually provides systems that can make better and faster decisions than humans. Since these systems will continue changing, we expect them to be able to transform our competence to exploit information at different scales.

4.2 *The Advantages of Using Machine Learning in Finance*

As we mentioned before ML is expanding to more and more sectors of our day to day, and finance is not an exception. Currently, ML is seen as a key resource in finances, including aspects such as asset management, risk level assessment, credit score calculation, and even loan approval (Israel et al., 2020). The industry of financial services usually deals with huge data volumes, related to daily transactions, vendors and customers or payments. As we have seen, this is an ideal environment for ML. This results in more optimal processes, lower risks, and better designed portfolios.

Some of services being improved by ML are:

- Algorithmic trading

The use of different types of algorithms to choose the best trading decisions is usually known as Algorithmic trading. Mathematical models to scan trading news and activities are developed by traders, with the goal of identifying the actual facts that affect prices (Jansen, 2020). In opposition to humans, these algorithms are capable of assessing vast amounts of data at the same time, and therefore make thousands of operations every day. Moreover, very often humans make decisions biased by their own emotions or feelings, which make them judge incorrectly. Algorithmic trading does not have this limitation. Different financial institutions employ algorithmic trading in order to automate their trading activities. Some of the most used algorithms are SVM and DNN (Abedin, 2021).

- Automated trading

Robo Advisor and Quant Advisor are both online investment platforms that provide an automated portfolio management service. That is, they offer a computerized advisory service based on algorithms without any human intervention in the investment decisions. Mostly they are based on NN and RL. They do not depend on the bias or fears that people may have. The main difference between these two fintech products is as follows:

- The Robo Advisor performs passive management through ETFs and index funds that try to replicate an index. Robo-advisors are online programs built using ML, which provide automatic advice to investors. The applications use algorithms to enact a financial portfolio depending on the investor risk tolerance and actual goals (Rosov, 2017).

  Robo-advisors tend to be cheaper than their human equivalent, requiring lower investments. Robo-advisors ask the investors to fill in their investment and goals, so they automatically find out which can be the highest returns or investment occasions.
- The Quant Advisor proposes an active management, looking for opportunities that the market generates at any moment of the economic cycle. They are based on quantitative analysis and are uncorrelated with the markets. Quant-advisors seek to provide tools that allow obtaining positive returns without the exposure

to the risk of volatility of the market being high. This way of investing avoids the high sensitivity of other methods to both bullish and bearish movements in the market. The instruments most used by quant-advisors tend to be listed futures, either on equity indices, on fixed income bonds, commodities, or currency.

- Fraud detection

Given the transactions from third parties, as well of the number of users doing these transactions, keeps on increasing, the protection against the security threats appearing in finance is becoming more and more and important. The losses from banking related frauds account for billions of dollars each year and therefore an efficient detection of these frauds is something the finance industry certainly looks forward to. One of the main sources of risk in this case arises from the fact that companies store most of their data online, which make security breaches easier to appear (Corporate finance institute, n.d.). Previous fraud detection systems were based on simple rules, which could be broken smoothly by many modern fraudsters. Therefore, ML is becoming the most common solution against this type of fraudulent financial transactions. The simplest solutions are created without the development of more capabilities, taking instead different data types to identify anomalies. More complex algorithms can provide data codes and/or scores that a real-time engine uses to decide. Mostly ML works by going through large data sets to detect deviations to be further investigated by a human expert. These algorithms are based on checking features of transactions, such as IP address, physical location, or the history of the account. This way the algorithm can find out if the transaction under scrutiny is compatible with the behavior of the holder of the account. Some ML algorithms even block the transaction entirely, if there is at least a 95% chance that it is fraud. ML algorithms need only seconds to evaluate a transaction. The identification of fraud is a binary classification which uses one of the most efficient binary classifiers, such as SVM, Decision Trees, or Logistic Regression. CNN and image identification are some of the other big improvements in fraud detection.

- Loan/Insurance underwriting

In the banking and insurance industry, ML is changing the way insurance companies and banks serve their users. Like many other industries, insurance and loans are data-driven. Companies access vast amounts of data from consumers, which can be later used to train ML algorithms that facilitate the underwriting process. ML algorithms are capable of deciding on underwriting and credit scoring, to help companies saving money and time. This means algorithms can be trained to assess consumer data, compare it to stored examples, look for similar cases, and finally decide if it is a good idea to accept a client for a loan or insurance. Examples of items these algorithms can use are the age, credit behavior, job or income.

Using ML, banks and insurance companies are realizing that the benefits are increasing, some of the most visible ones are:

- Increase in loyalty. Its use entails an increase in the speed of the response and the success in the resolution of incidents.

- Efficiency in claim procedures. They send people to the corresponding area or department to be attended quickly.
- Success in resolving incidents. ML makes an exhaustive analysis of the most successful claims and resolutions to offer the solution that best suits the client's needs. Correct management gives greater confidence in the insurance brand.
- Fraud prevention and detection. It is also used to calculate, based on a series of predictive algorithms, the probability of a customer trying to commit fraud in a claim, since it extracts data from internal and external sources.
- Offer customization. ML helps to predict better loss risk, which means cost savings and custom premiums.
- New ways of assigning policy prices. Obtaining quality information has become a much simpler and more accurate task. Sensors in vehicles, buildings and even the possession of smartwatches, allow insurers to be more exact in measuring and forecasting a risk. The information reveals data on the customer's lifestyles, driving habits, and so on. As a result, insurers can set prices without running risks and in turn allow consumers to purchase services according to their needs.

- Process automation

Process automation through ML appears in almost every field of knowledge nowadays. The reason for this is that the type of ML algorithms we have been discussing in this chapter allows the replacement of several manual tasks, avoiding repetition and helping to increase productivity. This topic will be further discussed in the following section *Machine Learning to automatize interface with final users*. The automation of processing in banking shows how ML can improve finances. Examples of these new applications are:

- Using NLP, a bank designed a contract intelligence program to deal with legal documents and automatically obtain data from them. If this were to be done manually, the reading of 12,000 business credit agreements would need around 360,000 h, compared to just a few hours using this type of ML tools.
- Another bank uses a chatbot controlled by ML with NLP through the Facebook Messenger tool to connect with users and aid with accounts and passwords.
- A Ukrainian bank made use of chatbot assistants on its online platforms. Chatbots accelerated the answer to customer inquiries of every type and decreased the need for human interaction.

Despite all the improvements ML can bring, even resource-rich companies often find it difficult to make the most out of this technology. It is not enough to have an adequate software infrastructure. Instead, it requires a clear idea, good technical skills, and courage to achieve worthwhile ML-related goals. Even so, many economists and financiers predict that within a few years a large part of financial processes will be developed through ML.

4.3 Industry 4.0 and Machine Learning

The goal of the so-called Industry 4.0 is to transform industry and make it intelligent. Thanks to the use of Cyber-Physical Systems, the Internet of Things (IoT), Cloud Computing, and Big Data; it is possible to collect, store, and compute a large amount of data (Datta & Davim, 2021). Due to the amount of data, manufacturing is one of the main industries that uses ML technologies to its fullest potential and, to achieve it, the industry needs to process all this data in an efficient way. In order to remove the manual data collection, the use of sensors was added. Some examples of these sensors are: humidity and temperature, cameras that allow recognition of shapes and objects, location and position sensors in space, pressure sensors, temperature sensors, etc. Another important part in this procedure is digital image processing. This technique allows the analysis and processing of digital images with computers to extract useful information from them. In manufacturing it can be applied to a lot of tasks. However, the application of ML algorithms is required to create models that give value to this data and facilitate decision-making.

There are many and diverse applications of ML in the industry such as:

- Production

In production, vision systems and robotics are combined with ML algorithms to improve processes and increase productivity. In fact, it is possible to automate tasks with such a variability that a traditional robot could not carry out on its own: recognizing and locating types of parts, processes and variable trajectories, etc (Monostori et al., 1996). For this reason, it allows in many cases to reduce costs and increase the competitiveness of companies. Moreover, improving sustainability is currently one of the challenges for manufacturers (López-Manuel, et al., 2020). In this regard, a full understanding of the processes, as well as finding simple ways to improve them, is a must. This is complicated in some cases, given how large the lines of equipment and processes can be, so it is crucial that manufacturers can figure out the best ways to implement these. For this, ML can provide interesting handles, such as NNs, which can be useful to obtain the most optimal methodology to carry out a process. The main advantage of this automatization is to find out which issues the production can face and how to tackle them. Some of the production issues that are specially solved by ML are:

- Failure rate reduction

They allow the detection of failures and their reduction, which has a direct impact on the quality of the production and its improvement. Mistakes from the past help to improve the process.

- Stock optimization

Stock optimization pursues one objective: maximize profit or minimize costs. With this clear, it is necessary to consider a series of restrictions that can influence costs (maximum stock, transport cost, delivery times, etc.). ML is usually applied to predict the sales of a product and hence predict the stock and a possible stock shortage. Also, the stock and sales prediction help to set the prices and offers for the products.

- Process automation

With these algorithms it is possible to automate processes that would not be possible without learning-based systems: variable inspections, changing environments, etc.

- Quality

Quality improvement investigates the relationship between the features of a product result and how well it matches its design and desired performance. The quality index can be objective (a physical or chemical test) or subjective (a human test). For this type of problems, the sensor's information is used by a NN, that can be either a DNN or a CNN to process images, to determine if the quality check is passed or not.

ML can be used to predict the quality of a product in two separate ways:

(a) Based on its design, modeling the result allows to simulate assorted products and adjust the design to improve quality.
(b) An anomaly detector algorithm can automatically indicate issues in the production line that can result in quality flaws. For example, to detect defects in parts, manufacturing surface defects, paint, etc. They also allow quality checks to be carried out in an assembly process, the presence or absence of parts, the inspection of welds, etc. These mistakes can be of different range but, in any case, have an overall influence in the production so that removing them in the initial stages would certainly help to save resources.

- Logistics

The supply chain generates a huge amount of both structured and unstructured data every day that can only be exploited thanks to artificial intelligence. Logistics is based on physical and digital networks that cannot be optimized by humans due to their high complexity (Kotsiopoulos et al., 2021). Therefore, the goal of ML is to transform reactive behaviors into proactive ones, manual into automatic, and standardized into personalized ones. For autonomous transport to be accepted, it must exceed the capabilities of a human behind the wheel, starting with the perception of the environment and its ability to predict changes in it. This is possible thanks to the combination of technologies that build a three-dimensional map of the environment. DN is responsible for processing data to identify traffic signs, detect obstacles, and other cars on the road, as well as comply with traffic laws. These algorithms are essential for the machine to acquire knowledge, since humans are not capable of programming all the possible situations that may occur on the road.

- Maintenance

In recent years, the cheaper sensors and their smaller size have facilitated the obtaining of valuable information on the state of the machines. Specifically, by measuring different points and characteristics of a machine, it is possible to have an almost real-time view of its status. By analyzing data obtained from different machines, models capable of predicting failures can be generated. Likewise, this helps to improve processes, avoid failures before machines break down, avoid production stoppages and reduce preventive maintenance time. Thanks to ML, models are created from this data in order to detect possible anomalies before they happen. This is called predictive maintenance, and its application is one of the most reliable ways to prevent machines from failing and damaging the production process (Chuprina, 2020). Although the concept of predictive maintenance is not new, the development of data collection, storage technologies and ML applications have contributed to create a new perspective on this term. In this way, the contribution of data from many sources is treated with more complex algorithms that have allowed a reduction in maintenance costs of between 10–15%. Although ML manages to analyze data and learn extremely effectively, it is important to highlight the role of humans when developing and improving tools related to maintenance. Both the experience of the workers and the historical data of the machines are basic information for the systems to work. In addition, feedback from an experienced person is needed to adjust the algorithms and validate the results they show. After all, decisions are made by humans based on predictions made by ML.

- Ergonomics

Regarding the working conditions in the production and assembly positions, Motion Analysis Systems (MAS) are used to create detailed reports on both the productive and ergonomic performance of the worker. This is achieved with a hardware called MOCAP (Motion Capture) integrated with software based on neural networks specialized in the treatment of images and videos. The adoption of MOCAP technologies in industry has grown in importance with the advent of smart factories. MOCAPs were originally designed to recognize movement in video games, but now the benefits of using the same principle to study and improve manual activities have been seen. Another way of applying Machine Learning in ergonomics is achieved with the sensors equipped in smartphones. These sensors provide information on the location and movements carried out by the user, such as the step counter. Unlike other devices, smartphones are cheap, easy to use and do not require a lot of maintenance.

- Security

The increased use of the Internet, both in social and work life, completely changes the way people learn and work. However, the number of cyberattacks is increasing in the same way. In the period of digital transformation in which companies find themselves, a technological incident of this type can put an end to the continuity of the company. For this reason, cybersecurity arises as a set of technologies and processes designed to protect networks, computers, programs, and data from possible attacks and threats. Powerful Machine and Deep Learning algorithms in cybersecurity are mainly used for malware analysis and intrusion detection and prevention. The development of these algorithms is driven by the need to anticipate a cyberattack and restrict access to infected files or programs.

- Product classification

Finally, we have a point that we can relate to some of the previous ones, such as quality control and artificial vision (Wuest et al., 2016). And it is that, both the sensors and the cameras help to identify aspects that will be decisive for the classification of the products based on the measured parameters.

The biggest companies in the world have been utilizing ML in manufacturing and investing millions in related developments. Some of them even have been digitizing the factories and buildings for many years. Although this concept sounds very avant-garde, the truth is that it has been in place for a long time. In fact, it is not the future of the sector, but the present. What ML proposes is to take the company to a much more advanced level of digitization.

4.4 Machine Learning to automatize interface with final users

Whether we want it or not, ML is behind the curtains every time we do a browser search, an online purchase or even take a picture with our phone. In the last decade ML has improved a lot our relations with machines and allow them to achieve more often what we want from them (Dudley and Kristensson, 2018). Of course, many of the advances in this area are very well protected and licensed by software companies and we do not intend here to disclose any of those. Instead, we will briefly go through some examples to show how ML affects the user interface, and how it relates to the different methodologies we are discussing in this chapter.

- Searching

Our favorite search engines have ML algorithms behind their success. We use the most popular engine, Google, to briefly explain the interaction between ML and internet searches. Google searches do not work with strings per se, but “entities”. Entities are unique items based on semantic analysis and they are formed when the strings in a search are grouped according to aspects such a context of the search, history, or ranking. Entities can be linked through relations that exist before any actual search is done. In the actual search, entities can be connected through contents. For instance, in the question “is Brazil beautiful?”, “Brazil” and “beautiful” are the entities and “is” the content. “Brazil” and “beautiful” were already connected entities through links given that is a statement that appears often!

Apart from entities, the other element in Google searches where ML plays an essential role is RankBrain. RankBrain is a system that connects different searches, where similar entities are involved, also accounting for additional information such as who is making the search, from which device or its location. Furthermore, RankBrain is continuously learning these connections through an iterative process, where it keeps improving to become more accurate. With this, it can provide a user optimized ranking whenever a search is performed. RankBrain runs in Tensor Processing Units (TPUs) (https://cloud.google.com/tpu/docs/tpus) which are application-specific integrated circuits developed specifically by Google for ML applications.

- E-mail filtering

E-mail filtering is usually given as one of the typical examples of ML application, especially around natural language processing. However, this is still an area where developments keep being made, given phishing, spam, or in general unwanted messages threaten our mailboxes increasingly often (Karim et al., 2019). Apart from annoying for the average user, spam mail causes millionaire economic losses across the world. An e-mail is composed of a TCP/IP header, a Simple Mail Transfer Protocol (SMTP) envelope, an SMTP header, and the body. All these elements can give us hints of whether an email is spam and are therefore tackled by the spam filter algorithms. The first obvious solution against spam does not involve ML per se but focuses on ensuring the sender/receiver of the mails are authenticated and authorized. For instance, the use of encrypted mails is becoming the norm. Other solutions involve *hashing* of e-mails or filtering based on regular expressions.

Focusing on ML, although both supervised and unsupervised solutions are possible, the latter remain the most accurate and frequent, with algorithms focusing on SMTP headers and body of emails. The most popular solutions are currently SVMs and Naïve Bayes, although Neural Networks are becoming more and more employed. For instance, CNNs can be used to detect spam based on images associated with the emails. As explained, this is an area that keeps improving and with a lot of work ahead: as spammers become increasingly sophisticated, so must be the methods to detect their mails.

- CAPTCHAs and cybersecurity

CAPTCHA stands for "Completely Automated Public Turing Test to Tell Computers and Humans Apart" and represents a particularly smart use of ML for user interface. CAPTCHAs were designed for cybersecurity purposes, as a barrier against hackers, to ensure online interactions are with humans and not with machines (Yang, 2018). We have all experienced CAPTCHAs when browsing online, but what many might not know is that by entering CAPTCHAs we are training a ML algorithm. The first CAPTCHAs were used to directly ML algorithms to "read" texts more accurately. More recently, they are becoming a way to train CNNs for image recognition. The fact that CAPTCHAs are becoming harder and harder for humans to solve is a sign of how hard it is becoming to find tasks that machines are not able to perform.

Other than helping with the defense, ML can also be used to perform cyberattacks. For instance, GANs have been shown to improve brute force for cracking passwords, based on the real passwords obtained from actual password leaks (Hitaj et al., 2019). Other possibilities include the use of RL to design evasive malware or RNNs for preparing advanced phishing attacks based on information retrieved online.

- Translation

An automatic and accurate translation between languages is one of the most ambitious targets of ML algorithms. This problem is usually targeted through RNNs, since these somehow emulate the reading process, keeping a memory of the processing steps, so that outputs from one step can serve as input for the same step at a later stage. An alternative is the use of CNNs, as in some commercial solutions, such as DeepL Translation.

Other recent developments are not directly involving CNNs or RNNs, but instead rely on novel NN architectures. For instance, a new network based on "attention mechanisms" (a type of encoder-decoder architecture) has been able to improve the scores achieved by human professionals in Czech–English translation of texts (Popel, 2020).

Two more areas of work are related to translation, although they are not as much progressed since they involve additional degrees of complication. The first is speech recognition and translation, which would help to achieve online translation of conversations. The second is augmented translation of texts, such as translating a written panel live just by pointing our phone cameras to it. Both are combinations of different ML techniques. The first layer involves the speech or image recognition, explained later in this subchapter, and the second the translation itself, as discussed here. In particular, the online translation of conversations is especially challenging and rewarding at the same time, so a lot of developments are expected in this area in the future.

- Advertising

We have all experienced it: doing an internet search for a product, and then spending several days watching ads for related products even when you are not interested anymore. Private companies keep learning increasingly about us, and this information can be more effective in tempting us to buy what they have to offer. ML is excellent at deciding which is the best way to trick us, building publicity based on our browsing history, interests, or even traces of our personality companies has been able to learn (Hwang, 2019). Moreover, the ability of ML to find hidden patterns is especially useful in this context, connecting products with customers that would have not been otherwise targeted.

This can be taken one step further, by finding out how certain images can be more appealing on marketing depending on the viewer characteristics or even designing the right ad for the right person. For instance, a study has designed a model that is trained to tailor personalized ads based on simple LASSO regression.

Another possibility is the use of so-called "contextual relevance", based on natural language processing and DL. In this case, ads are designed to be displayed only in certain contexts. For instance, this can be used as an aggressive strategy to have your ads placed next to those of your competence. A final example involves ultra-personalized publicity. This involves creating different versions of ads to be shown based on who is watching them and in which context (i.e., next to a video, in a newspaper website). This approach targets not only our personalities but also our moods when we are reading or watching specific types of content.

- Customer support

The use of ML in customer support revolves around speech recognition and natural language processing. Companies are starting to use ML to provide a quick and direct way to most of the issues customers raise, as well as to save costs by reducing the need for human intervention. The huge amount of client interactions that big companies must deal with, while a burden, is also interesting from a purely ML point of view, since it generates data for ML algorithms to be trained.

The first line of attack for ML in customer support is the automatic identification of customer issues. The same issue can be phrased in an almost infinite number of ways, and being able to recognize this similarity is an important challenge for natural language processing algorithms. An additional input to recognize issues might come from "social listening", which relates to marketing and refers to the tracking of online activities that users have performed prior to their queries. A variation of this task is the assignment of agents to customers. In that case, the idea is not so much providing a solution to the issue but identifying the topic of the inquiry to find the right human expert. This is becoming more frequent in call centers, which, as with translation, adds an additional layer of complication through speech recognition. In this case, not only *what* is being said is of interest, but also *how*. For instance, measuring the volume or pitch of the voice can be used to determine the priority of the call: companies want to listen first to angry clients!

A more ambitious approach is that of the famous chatbots. Those intend to provide a full service to customers (intend-based chats), even attempting to establish conversations (flow-based chats). Chatbots turn out to be one of the most demanding applications of natural language processing but is an area where a lot of research is being done and significant improvements are expected in the near future.

- Speech recognition

We have already seen in this chapter how speech recognition has multiple applications in the user interface. From translation to customer support, speech recognition is a pillar in which many ML applications stand. Other important examples involve virtual assistants (such as Alexa or Siri), enhanced biometry using voice recognition, transcription of meetings or interviews, language learning, or automatic subtitling of contents (Kamath et al., 2019).

The first step for speech recognition is sampling, i.e., taking pieces of the audio input and digitizing them. Sampling usually also involves a mathematical operation called "Fourier transform" which helps to disentangle the frequencies involved. Note sampling is a classical problem, but it has lately become a focus of ML algorithms themselves, such as CNNs.

The next step is the conversion of the sampled input in the actual text outcome. For this, RNNs are usually employed since, as discussed above, they are good at tasks related to "reading". In particular, the so-called long short-term memories RNNs (LSTMs), specialized in time series data, are frequently part of this processing step.

Large datasets of written texts are an important additional source of information here, which can help to better interpret what comes out of the speech. For instance, if the network doubts between “apple” and “attle”, this will give a much higher score to the first one, given it is a word that appears more often in English.

As in other cases in ML, speech recognition models are constantly “learning” from us. Note every gadget now has virtual assistants whose interaction with us is constantly used to improve their response through training!

The final challenge is finding end-to-end models that include all these intermediate steps in a single trainable algorithm. An exceedingly popular solution, applied both in natural language processing alone, but also in speech recognition, is Connectionist Temporal Classification, which provides a sequence of continuous output probability labels.

- Image recognition and manipulation

Image recognition is also present in many aspects of our daily lives, such as when social networks recognize our faces in pictures. Increased instances of use appear every day to make our interaction with machines easier. Recent examples involve the recognition of artworks or the visual search for products for purchase.

The ML analysis of images can involve classification, detection, and recognition (Voulodimos et al. 2018). These are obviously correlated, but not the same. The first, classification, is about categorizing images, such as grouping all pictures where a firefighter is located, among a group of pictures with different professionals on them. The second, detection, is about locating specific elements in an image, for instance, finding out if there is any dog in a picture with several components on it. The third, recognition, is somehow a combination of the other two, and becomes closer to what a human can do. For instance, it would imply finding out if a particular person, with a known face, is present in a photograph.

All three of classification, detection, and recognition heavily rely on CNNs. As with speech recognition, the preprocessing of the images is crucial to begin with, with decoding and resizing as frequent additional steps. Then, the processed image is fed into a DNN for the required task. This will usually use image patterns to achieve its goals. Several pre-trained models, such as AlexNet, ResNet, SqueezeNet, or DenseNet, exist in the market, making the analyses easier. Models for specific tasks can make use of transfer learning to make the training quicker and more reliable.

Another recent and related late development is that of manipulation of images or video through DNNs. When this manipulation involves humans, this results in the popular “Deep fakes”. These have had applications in films or social media, but also fraudulent ones, such as blackmail or fake news. From a formal point of view, the manipulation of images involves the use of autoencoders or GANs. The first encoded images into a latent space, which can be then decoded through a pre-trained model to imitate, e.g., someone else’s characteristics. As for GANs, a first network generates fake images that a second network tries to discern from real ones, creating a zero-sum game that provides increasingly perfect results.

- Transportation: dynamic pricing and routing

Many of the advances regarding transportation happening in the last years are also related to the development of new ML models (Fitzek et al., 2020). This goes from the fluctuations in the price of plane tickets to the choice of the best route to go from one place to another in our car.

As for the price fluctuations, they arise from a well-known law of the market that the price of items sold will increase if there is more demand for them. This principle can be extended to other aspects important for flag carriers, such as competitor prices, price of fuel, market movements, or even the habits of customers.

Dynamic pricing consists of building models that consider all these aspects to decide the best fares at any moment. The same principles used in plane tickets also apply to ride-hailing, hotel bookings, or in general e-commerce. Even if it is possible to establish rule-based systems for dynamic pricing, the best solutions come, as one could expect, from ML models. As we are often seeing in this chapter, ML algorithms can find hidden patterns and determine the best pricing strategy for a company. LSTMs are the most frequent type of models for dynamic pricing.

The custom models designed by companies provide a great degree of flexibility, since pricing can be modified according to expectations from customers and the allowed margin of profit that the company might have at a specific moment. This can be all done real-time, with no need for manual adjustments!

Then, concerning routing, the existing algorithms, such as Google Maps, are able to provide increasingly accurate answers about aspects like best route or time of arrival. For this, Google Maps feed several types of NNs with information regarding the type of road (highway or not, straight or curved, quality) as well as, of course, the traffic flows. For this, Google uses the so-called "Supersegments", which are ensembles of roads that share their traffic volume. Then, to estimate the time needed to transverse each Supersegment uses a type of NN called "Graph Neural Network" (https://deepmind.com/blog/article/traffic-prediction-with-advanced-graph-neural-networks). These extend CNNs and RNNs by including the concept of proximity, loading the network of roads into a graph with connections and edges (Fig. 12). This type of NN has been able to improve the accuracy in the time of arrival provided by Google maps up to 50% in some places!

Fig. 12 (Left) Example of CAPTCHA. CAPTCHAs are used in cybersecurity as way to distinguish humans from machines, but they are also useful to train ML models. (Right) Example of a "Deep fake" image, generated using adversarial learning. Free images to use a prior [https://en.wikipedia.org/wiki/File:Unexpected_CAPTCHA_encountered.png] and [https://commons.wikimedia.org/wiki/File:Sw-face-swap.png]

4.5 How Machine Learning is revolutionizing Medicine

The use of ML in medicine is becoming increasingly popular, and the possibility to save lives makes any progress in this area particularly attractive (Rajkomar et al., 2019; Goecks et al., 2020). The scalability of ML algorithms makes those especially useful for medical applications, and an innovative way to make the most of some of the vast existing datasets. The fact that in ML a model, rather than having to follow strict rules, learns from examples can make these models very flexible, going beyond many of the existing traditional solutions. In this regard, ML algorithms are normally able to find hidden patterns in data, which might be challenging for health professionals. Finally, the fact that ML algorithms can be run instantaneously, once they have been trained, are a key aspect for certain applications, when a fast decision might be the key to avoid permanent harm.

Now we discuss the current trends of ML in medicine, providing specific examples in diagnostic, treatment, and health management.

- Diagnostic

ML has been shown to perform as well as medical professionals across different specialties in terms of diagnostic, and a joint human–computer diagnostic is currently regarded as the optimal solution for the future. Computers would offer potential diagnostic options, and it would be left to the health care professional to make the final decisions on how to proceed. Examples of use of ML in diagnostics are:

- Imaging: the use of CNNs is becoming very common in computer vision applications, e.g., cancer diagnostic, Optical Coherence Tomography (OCT) in eye conditions such as diabetic retinopathy (DR), radiological images or magnetic resonances (Esteva et al., 2019). As explained above, the use of ML can be sometimes especially useful when an urgent diagnostic is required, such as radiological images of the brain, when patients have limited time before permanent damage occurs. A remarkable alternative to CNNs around medical imaging is Deep Belief

Networks, made by accumulating autoencoders, which have been applied, e.g., to magnetic resonance images to detect signs of schizophrenia.

- Molecular tests: DL can be used for the interpretation of genetic data, for instance, for phenotype prediction, or in genome-wide association studies, making use of the ability of this type of algorithms to make complex associations that are hard to be found for humans.

- Treatment and prognosis

ML algorithms are excellent at accounting for multifactorial effects, overcoming the limitations of practitioners. This is often crucial to decide the best treatments to follow, or in general, to generate suggested therapies for experts to choose. Moreover, the capacity of ML to learn patterns from existing databases can be also useful to provide predictions about the evolution of patients or, more specifically, about the outcome of a disease.

A new remarkable use of ML in medicine has led to the discovery of new antibiotics (Stokes et al., 2020). For this a DNN was trained to predict how well different molecules inhibited the growth of Escherichia coli using a collection. Then, this network was applied to different chemical libraries, comprising hundreds of millions of molecules, to identify those with potential as antibiotics. The best candidates were then tested in mice, confirming this potential. Another powerful example is the use of RL for robotic-assisted surgery, where the RL algorithms learn from the motion of human surgeons.

- Health management

The most promising area for using ML in health management revolves around electronic health records. The reason for this is that ML is often useful to deal with massive amounts of data that are exceedingly difficult to interpret coherently as a whole (Ngiam and Khor, 2019). Apart from this, natural language processing tools, such as RNNs, are starting to be used to help in the analysis of the texts in, e.g., medical reports. The same goes with speech recognition techniques to transcribe conversations with patients. Another interesting example are unsupervised learning algorithms, which might be particularly useful here. For instance, tools such as autoencoders can learn representations to then reconstruct unlabeled data and perform diagnoses.

Other than this, wearable devices can be used to generate data for ML applications. This data can be used in health management to provide real-time health-care suggestions. The use of smartphones as an additional source of health information can be immensely powerful in this regard.

The use of ML tools in medicine is a rapidly moving field. Notably, in 2018 the American Food and Drug Administration (FDA) approved for the first time a ML-based DR diagnostic system (Abràmoff et al., 2018). This is significant since DR is one of the leading causes of blindness in the world. This automatic system, using CNNs to interpret OCT images, achieved a sensitivity of 87.2% and a specificity of

Fig. 13 (Left) Magnetic resonance of a human brain. CNNs are capable of finding patterns that might be invisible even to the expert eye, and can do so much faster. (Right) Electronic Health record. EHRs are becoming ubiquitous in medicine, and their use is the great interest in ML to facilitate the access to medical data. Free image to use a priori [https://pixabay.com/service/license/]

90.7% when compared to the standard method used so far, using human readings of the OCT images.

This progress in the use of ML in medicine has inherent challenges, some specific to this area, related to the management of data and ethics. For the first, the main issue is the collection of formatted, unbiased and uniformized data. A workaround for this comes from ML itself, using data curation, i.e., algorithms that can pre-classify data well enough to make ML algorithms more performant. Moreover, concerning the bias, this might arise from the use of data produced in merchandised health systems. In those, for instance, sometimes unnecessary care is given, while, on the other hand, data from certain population segments, such as those with no insurance, is not available. Finally, for ethics, some of the issues to be addressed are the privacy, security, and control of patients' data, given biomedical information might be especially sensitive, so hard to share. A novel solution to this is federated learning, where just the algorithms have full access to data that is split in several independent nodes.

ML should in general not be seen as a replacement to human medicine practitioners, but rather to assist them. Instead of having them as black boxes, it is important that practitioners and patients understand how ML algorithms do their predictions. This might be crucial, for instance, to establish the liability in cases of medical errors (Fig. 13).

4.6 *The Rise of Machine Learning to Accelerate Science*

Although ML algorithms seem to naturally fit in the category of "applications", their use in natural science has exploded in the last few years. The capacity of ML for interpreting data and finding hidden patterns can provide a lot of advantages to understand the Nature around us. One of the most interesting aspects when using ML in natural science is the availability, in some cases, of exceptionally large datasets.

Spectacular examples are those of colliders in particle physics, producing millions of events per second, or the catalogues of molecules in chemistry, with hundreds of millions of samples. We next show some examples of the use of ML in science, focusing on biology, chemistry, physics, and geoscience.

- Biology

The increase in the availability of large datasets from biological systems in the last few years has accordingly led to incredibly significant expansion in the use of ML in this field (Tang et al., 2018). These datasets comprise aspects such as molecular variables, genetic variation, or microbiome composition. As in other areas, ML is especially good at deriving nonlinear relations in biology datasets that would otherwise remain hidden. The main applications of ML in biology appear in genetics and biochemistry, such as for genome annotation or the study of metabolic functions. However, other priori simple aspects, such as the study of cells, have also benefited from ML. For instance, the free tool Cell Profiler (https://cellprofiler.org/), which was developed to measure cell properties, has recently incorporated DL to record and combine features and produce new features for cell analysis.

As mentioned, ML is starting to be one of the main tools in genetics. For instance, since RNNs are particularly good at dealing with sequential data, they are becoming key in the understanding of DNA arrays or genomics sequences. RNNs, in combination with CNNs, have been used in genome-wide analysis or to predict gene expression. Likewise, other DNNs have been employed to build predictive models of RNA to identify potentially pathogenic mutations. Another interesting example is the prediction of the structure of expressed bioproteins through autoencoders.

Biological networks are also a good target for ML algorithms, given they involve large multi-dimensional datasets (Camacho et al., 2018). The study of these networks relates to the interactions between cellular systems through biomolecules, which are responsible for the structure and behavior of living cells. For example, in plant studies, SVMs have been used to study pre-microRNAs and mature microRNAs, which help to provide a better defense against pathogens. In the same regard, SVMs have also been applied for the development of disease-resistant plant varieties.

As for disease biology, DNNs are being employed to learn the key features that allow the recognition of healthy states, identifying the interactions and biomolecules that define these. A similar principle can be applied to the discovery of new drugs, through the prediction of drug toxicity for tackling cancer cells. These developments make use of the advantage of ML algorithms when dealing with multi-label datasets. As a final significant example in this area, transfer learning from other well-studied microorganisms has been used to help understanding the human microbiome, composed of the microorganisms inside the human body.

ML has exciting potential for other future applications in biology. A remarkable example is the use of DNNs for creating synthetic gene networks from biomolecules so that these would serve as "circuits" to modify the cells behavior, acquiring new capabilities of interest. For this progress to happen, the main challenges are achieving large enough biological datasets for training new ML algorithms and the interpretation of the consequent ML models to help understanding the underlying biological mechanisms that are at play.

- Chemistry

Chemistry is no exception in the surge of ML in fundamental science. From chemical sensing or the design of experiments to the discovery of new materials or molecules, the use of ML algorithms is becoming increasingly frequent by chemists (Rodrigues Jr et al., 2019; Cova & Canelas Pais, 2019). The recurrent presence of patterns in Chemistry (in, e.g., combinations of functional groups or crystalline structures) is of special interest here, since this can be exploited by ML algorithms to find hidden properties. The source of the chemistry datasets needed by ML algorithms is varied: this might come from computer simulations, diverse types of sensors, specific experiments, or public datasets (such as in crystallography or materials data).

The discovery of new materials can be hard and expensive through traditional methods and is therefore a good target for the use of ML algorithms. In this case, it is essential to have at hand software to make proper simulations of materials and databases with their properties, together with an adequate representation of the relevant information. An example of this is the SMILES representation, which encodes molecular structures, and which has turned out to be especially useful when used in association with different ML tools.

For instance, autoencoders can be used to convert SMILES (Weininger & Smiles, 1988) representations in a "continuous" space, allowing to obtain the most optimal chemical properties. Similarly, RNNs can be also used in association with SMILES representations to generate new molecules with better properties. An original development in this area, which is very correlated to ML, is the so-called "genetic algorithms". In this case, compositional and structural features of materials are interpreted as "genes". The domains of these genes are then scanned to search for the adequate "phenotypes", with the goal of achieving the target properties.

Another interesting approach in this area is the use of NNs or DNNs in synthesis prediction, with the network being trained to identify the sequence that is most likely to produce a specific compound and the optimal conditions for this to happen. In the same regard, RL algorithms have also been used for synthesis planning. Finally, crucial progress in this area is expected to come from computer-aided drug design. This might especially be important for the discovery of drugs tackling neglected diseases. Here, there is an obvious overlap with biology and medicine, described above. Advanced methodologies, such as graph-based structural signatures, DNNs or self-organizing maps (a type of NN) have already started to be used for drug design. These are especially useful, e.g., to predict the activity of bioactive molecules or minimize the toxicity of drugs.

Not surprisingly, ML has also appeared as a valid solution in quantum chemistry. These applications are correlated with some appearing physics, which we will be reviewing next. ML can supplement or directly replace complex calculations in quantum chemistry and predict quantities such as spin angular momenta or bond energies, as well as more complex aspects such as modeling electronic quantum transport. Moreover, the combination of density-functional theory (DFT), which is a quantum mechanical method widely used in calculations to model the structure of matter, with ML has turned out to be very efficient. For instance, DNNs or even more simple methods, such as random forests, manage to significantly reduce the computation time when dealing with DFTs. However, one of the main challenges that ML faces in this area, when compared to, e.g., computer aided drug design, is the lack of abundant training datasets.

Developing new materials for chemical sensing is becoming easier thanks to ML. The easiest example is the use of different classifiers, such as Decision Trees or kNNs, to achieve simple quality control for distinct types of compounds in medicine, food industry, or cosmetics. In the same regard, in agriculture ML can be used for disease control, to differentiate samples that have been affected by disorders. Finally, more advanced techniques with immense potential could involve microfluidic sensors to detect gluten in food or even its taste. For this, techniques such as, Random Subspace Methods, based on ensemble learning are especially promising. As for biology, the future of chemistry seems to be intricately linked to that of ML. Progresses with chemistry or bio sensors will be essential to provide the large datasets that ML algorithms require. Examples of areas where developments are expected and which were not covered here are analytical chemistry or catalysis.

- Physics

From a more fundamental perspective, tackling the statistical side of ML, to a more applied one, looking into how ML algorithms could be run in future quantum computers, ML is becoming omnipresent in physics (Carleo et al., 2019). The initial relation between ML and physics comes from statistical mechanics, given some of ML algorithms, such as Boltzmann machines, were an application of physics concepts to data science. Currently, the understanding of different ML algorithms, such PCAs or even NNs from a physics perspective, is an active field. This is important to fight against the usual view of methods in ML as black boxes and help interpret the data connections one can learn when running these algorithms. Conversely, ML also provides interesting handles to study physics systems, such as the use of RNNs in non-linear dynamical systems.

The application of ML has experienced a boom in particle physics, astrophysics, and cosmology in the last decade. The first example involves the use of BDTs for classification and regression in particle physics. This has lately evolved into other methods, such as CNNs in classification of particle showers (jets) increased trigger selections rely now on ML. For instance, the LHCb experiment at CERN has developed several of these selections, such as one that relies on the use of NNs specialized in quick evaluation in real-time environments (Benson et al., 2019). As for cosmology, methods based on NNs and BDTs have been developed to measure the photometric redshift and CNNs to estimate several cosmological parameters based on different dark matter measurements. As a final example, different ML methods, such as GANs, are also being currently used to denoise and deconvolute reconstructed parameters in particle physics and cosmology.

ML is also suitable to deal with the interaction of multi-body quantum states. A recent remarkable development in that regard is that of Neural-network quantum states (NQS), which are a NN representation of the many-body quantum wave-function (Carleo & Troyer, 2017). NQS are useful both for supervised and unsupervised tasks in the analysis of quantum systems. Another important associated aspect is the interrelation between ML and quantum computing. Apart from how quantum computers could enhance the training and application of ML algorithms, ML itself is also promising for the development of quantum computers. For instance, the tomography of quantum states, crucial in quantum information, can be done through deep learning approaches. Similarly, the preparation of qubits of control could be also improved via RL.

Apart from this all, there are several research areas in physics where ML is starting to be widely used, such as optics, climate science, or even in searches for exoplanets. Another remarkable area of research in physics is the use of new hardware platforms for ML application. An example is that of optical processing units.

- Geoscience

Geoscience is also starting to widely adopt ML techniques. Fields such as seismology or geomorphology are already using different variants of ML algorithms, and this is expected to increase in the future (Dramsch, 2020). The first uses of ML in geoscience involved, e.g., the use of SVMs for automatic seismic interpretation or of random forests for seismological applications such as event classification or localization in volcanic tremors.

As in other cases, DL is beginning to dominate the field of geoscience. Apart from their application in areas where other methods had been used before, such as seismic interpretation, this had led to a significant increase of applications. A powerful example is that of semantic segmentation for fault interpretation and salt detection.

More lately, U-nets, a type of CNN, have been used to interpret satellite data, to then prevent landslides or predict the arrival of earthquakes. Similarly, LSTMs, introduced above, are being applied to monitor volcanic activity.

Finally, although still at early stages, GANs are being applied with seismic data, for instance to help generate seismograms (Fig. 14).

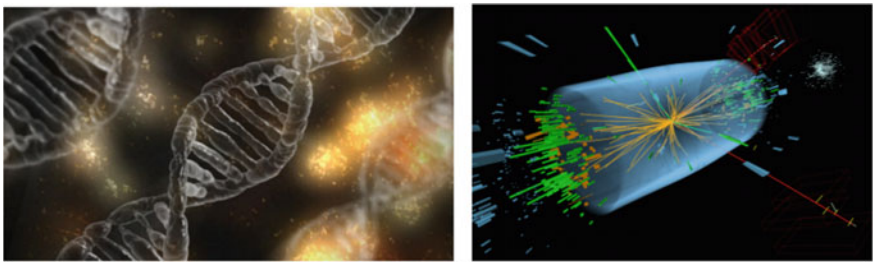


Fig. 14 (Left) DNA helix. RNNs are good at dealing with DNA sequences, helping in genome-wide analyses. (Right) Particle collision at the Large Hadron Collider at CERN. BDT and DNNs are used to classify and help reconstruct the particles produced after the collision. Left, free image to use a priori [https://pixabay.com/service/license/]. Right image, image courtesy of CERN [CC-BY-SA-4.0 license http://cds.cern.ch/record/1606502]

References

Abedin, M., Hassan, M. K., Hájek, P., & Uddin, M.M. (2021). *The essentials of machine learning in finance and accounting*. Routledge.

Abràmoff, M. D. et al. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy inprimary care offices. *NPJ Digital Medicine, 1*, 39.

Aguilar, V., Wu, H., & Montgomery, J. (2020). Machine Learning in Foundation Design and More. *TRB Centennial Circular, 103.*

Alake, R. (2020). Deep Learning: GoogLeNet Explained. Available at: https://towardsdatascience.com/deep-learning-googlenet-explained-de8861c82765

Amezquita-Sanchez, J., Valtierra-Rodriguez, M., & Adeli, Hojjat. (2020). Machine learning in structural engineering. *Scientia Iranica, 27*, 2645–2656.

Benson, S., et al. (2019). Real-time discrimination of photon pairs using ML at the LHC. *SciPost Phys., 7*, 062.

Brink, H., Richards, J., & Fetherolf, M. (2016). *Real-world machine learning*. Simon and Schuster.

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell, 173*(7), 1581–1592.

Carleo, G., & Troyer, M. (2017). Solving the quantum many-body problem with artificial neural networks. *Science, 355*, 602.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., ... & Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics, 91*(4), 045002.

CellProfiller (n.d.). Cell image analysis software. Retrieved in October 2021 from https://cellprofiler.org/

Chuprina, R. (2020). AI and Machine Learning in Manufacturing: The Complete Guide. Available at: https://spd.group/machine-learning/ai-and-ml-inmanufacturing-industry/

Cova, T. F., & Pais, A. A. (2019). Deep learning for deep chemistry: optimizing the prediction of chemical patterns. *Frontiers in chemistry, 7*, 809.

Corporate finance institute (n.d.). What is Machine Learning (in Finance)? Retrieved in October 2021 from https://corporatefinanceinstitute.com/resources/knowledge/other/machine-learning-in-finance/

Datta, S., & Davim, J. (2021). Machine learning in industry. Springer.

Dertat, A. (2017). Available at: Applied Deep Learning - Part 3: Autoencoders. https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

Fitzek, F., Granelli, F., & Seeling, P. (Eds.). (2020). *Computing in Communication Networks: From Theory to Practice.* Academic Press.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Glen, S. (2019). Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply. Available at: https://www.datasciencecentral.com/profiles/blogs/decision-tree-vs-random-forest-vs-boosted-trees-explained

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Google (n.d.). Cloud Tensor Processing Units (TPUs). Retrieved in October 2021 from https://cloud.google.com/tpu/docs/tpus

Gupta, P. (2017). Decision Trees in Machine Learning. Available at: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

Hasan, S., Shamsuddin, S. M., & Lopes, N. (2014). Machine learning big data framework and analytics for big data problems. *International Journal of Advances in Soft Computing and its Applications, 6*(2).

Hitaj, B., Gasti, P., Ateniese, G., & Perez Cruz, F. (2019). Passgan: A deep learning approach for password guessing. *International Conference on Applied Cryptography and Network*.

Hulten, G. (2019). *Building Intelligent Systems: A Guide to Machine Learning Engineering*. Apress.

IBM Cloud education (October 20th, 2020). Convolutional Neural Networks. Available at: https://www.ibm.com/cloud/learn/convolutional-neural-networks

Israel, R., Kelly, B., & Moskowitz, T. (2020). Can machines 'Learn' finance? *Journal of Investment Management*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3624052

Jansen, S. (2020). *Machine Learning for Algorithmic Trading: Predictive models to extract signals from market and alternative data for systematic trading strategies with Python*. Packt Publishing Ltd.

Kamath, U., Liu, J., & Whitaker, J. (2019). Deep learning for NLP and speech recognition (Vol. 84). Springer.

Kotsiopoulos, T., Sarigiannidis, P., Ioannidis, D., & Tzovaras, D. (2021). Machine learning and deep learning in smart manufacturing: the smart grid paradigm. *Computer Science Review, 40*, 100341.

Lange, O., & Perez, L. (2020). Traffic prediction with advanced Graph Neural Networks. Available at: https://deepmind.com/blog/article/traffic-prediction-with-advanced-graph-neural-networks

López-Manuel, L., León-Mateos, F., & Sartal, A. (2020). Closing Loops, Easing Strains: Industry 4.0's Potential for Overcoming Challenges of Circularity in Manufacturing Environments. In Circular Economy for the Management of Operations (pp. 23-47). CRC Press.

Monostori, L., Markus, A., Van Brussel, H., & WestkAmpfer E. (1996). Machine Learning Approaches to Manufacturing. *CIRP Annals, 45*(2), 675-712.

Osiński, B., & Budek, K. (2018). What is reinforcement learning? The complete guide. Available at: https://deepsense.ai/what-is-reinforcement-learning-thecomplete-guide/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830

Popel, M. et al. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparableto human professionals. *Nature Communications, 11*.

Ricquebourg, V., Menga, D., Durand, D., Marhic, B., Delahoche, L., & Logé, C. (2007). The smart home concept: Our immediate future. In *2006 1st IEEE international conference on E-learning in industrial electronics* (pp. 23–28). ICELIE.

Rodrigues Jr J. F., Florea, L., De Oliveira, M. C., Diamond, D., & Oliveira Jr, O. N. (2019). A survey on big data and machine learning for chemistry. Available at: arXiv:1904.10370.

Rosov, S. (2017). Machine Learning, Artificial Intelligence, and Robo-Advisers: The Future of Finance? Available at: https://blogs.cfainstitute.org/marketintegrity/2017/07/12/machine-learning-artificial-intelligence-and-robo-advisers-the-future-of-finance/

Scikit-learn (n.d.). 1.4. Support Vector Machines. Retrieved in October 2021 from https://scikit-learn.org/stable/modules/svm.html

Smith, R. (2020). How Machine Learning Will Transform Engineering. Available at: https://becominghuman.ai/how-machine-learning-will-transformengineering-e239e5774408

Sören Dramsch, J. (2020). 70 years of machine learning in geoscience in review. *Advances in Geophysics, 61*, 1-55,

Stokes, J. M., et al. (2020). A deep learning approach to antibiotic discovery. *Cell, 180*(4), 688-702.e13.

Swaminathan, S. (2018). Logistic Regression — Detailed Overview. Available at: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Tang, B., Pan, Z., Yin, K., & Khateeb, A. (2019). Recent advances of deep learning in bioinformatics and computational biology. *Frontiers in Genetics, 10*, 214.

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Science, 28*, 31–36

Wuest, T., Weimer, D., Irgens, C., & Thoben, K.-D. (2016). Machine learning in manufacturing: Advantages, challenges, and applications. *Production & Manufacturing Research, 4*(1), 23–45.

Yang, Y., & Yan, J. (2018). *Hands-On Data Science with Anaconda: Utilize the right mix of tools to create high-performance data science applications.* Packt Publishing Ltd.

Decision Support System Based on Deep Learning for Improving the Quality Control Task of Rifles: A Case Study in Industry 4.0

Luca Romeo, Riccardo Rosati, and Emanuele Frontoni

The quality control (QC) procedure was usually constrained by the high demands in time and resources, as well as a limitation in performance mainly due to the intra- and inter-operator variability and the risk of reproducing unwanted bias. Accordingly, the increasing amount of collected data open the realm of possibilities for designing and implementing a Decision Support System (DSS) empowered by Deep Learning algorithms for solving the QC task by overcoming these challenges. The work proposes a Deep Learning model for predicting the aesthetic quality classification (QC task) of rifles based on the analysis of wood grains. The task and the collected dataset originated from a collaboration with an industrial company. The higher performance (up to 0.86 of F1 score) by the proposed VGG-16 based model and the validation with respect to human annotator suggest how the proposed approach represents a solution to automate the whole QC procedure in a challenging industrial case scenario. The proposed solution allows to (i) speed up the QC procedure, (ii) minimize intra/inter-operator variability, and (iii) detect and mitigate unwanted bias by forcing the network to learn the features that correctly describe shot quality, rather than other confounding geometric features.

1 Introduction

In the context of Industry 4.0, the increasing availability of data, the advancements in computer power and breakthroughs in algorithm development have led Artificial Intelligence (AI) methodologies viable solutions in different Industry 4.0 domains

---

L. Romeo (✉) · R. Rosati · E. Frontoni
Department of Information Engineering, Università Politecnica Delle Marche, Ancona, Italy
e-mail: l.romeo@univpm.it

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8_3

such as predictive maintenance (Calabrese et al., 2020; Sakib & Wuest, 2018), decision support system (Romeo et al., 2020), production planning (Cadavid et al., 2020) and quality control (Escobar & Morales-Menendez, 2018; Wang et al., 2018). In particular, a growing area in this scenario is the QC phase, which is a fundamental step for detecting production issues and for classifying the compliance of the finished product. The application of Machine Learning (ML) and Deep Learning (DL) techniques offers great opportunities to automatize the overall QC process, saving time and resources and maximizing the performances, while easily generalizing in different contexts. In evidence of this, these methodologies for QC task have been employed in different domains, such as fabric and textile industry (Fu et al., 2019; Lee et al., 2018; Li et al., 2017), printing industry (Villalba-Diez et al., 2019), smart factory prototype (Ozdemir & Koc, 2019), laser-based additive manufacturing (Francis & Bian, 2019), welds mass production (Muniategui et al., 2019) and automotive industry (Iqbal et al., 2019; Peres et al., 2019). All of these solutions focus on quantitative and deterministic analyses: the dimensional control, the inspection of the material roughness, the patterned fabric defect detection and the test of production parameters are all measurable evaluation procedures. However, there is still no software tool which allows the automatization of all those qualitative QC analyses that are strictly human dependent, such as the aesthetic evaluation of a material.

For this reason, the introduction of a DL methodology in an undiscovered and challenging QC application was proposed: in this work the aim is to provide the aesthetic quality classification of rifles stocks by analysing the wood grains in a fully-automated manner. However, the strengths of DL methods in this scenario are limited to the inability to detect and reproduce unwanted bias in the prediction of the DL model.

Starting from these considerations and originated from a specific company's demand, the aim of the paper is to design and develop a DL model to (i) predict the aesthetic quality classification of rifles by analysing the wood grains and (ii) detect and mitigate unwanted bias in the dataset and in the prediction of DL model.

The work has the following main contributions in the Industry 4.0 scenario:

- the data collection and data annotation procedure of a real dataset specifically built to predict the aesthetic quality classification of shotguns by analysing the wood grains;
- the proposed DL model based on ordinal categorical cross-entropy loss and VGG-16 architecture for solving an undiscovered task namely the QC classification task;
- a thorough analysis for discovering and mitigating unwanted bias factors;
- the validation of the presented model with respect to the performance of a human annotator and the integration in a decision support system.

The paper is organized as follows: Sect. 2 provides an identification of the industrial use case scenario by showing the specific sector/company and their related main activities; Sect. 3 shows the task we aim to solve in this work, which is the QC of wooden stocks; Sect. 4 explains how this problem was solved before using AI solution; Sect. 5 gives details on the presented model, by describing the dataset and the DL method used. In Sect. 6, a comparative evaluation of our approach, bias mitigation and the validation by human annotator are reported. Finally, in Sect. 7, discussion and conclusions with future directions for this research topic are drawn.

2 Industrial Use Case Scenario

Benelli Armi Spa is an Italian firearms manufacturer based in Urbino (PU). Founded in 1967, it specializes in the production of semiautomatic rifles: the company produces more than 200 thousand weapons per year, exporting to 78 countries around the world (first market United States, followed by Italy and Canada). The weapons produced are mainly intended for hunting and sports activities, only a small part for military use or collectors. The production is completely made in Italy, except for the wood supply that is imported from American walnut plantations and, in a small part, from Turkey and Pakistan.

The key factor that allows Benelli to be a living and constantly growing company is the investment in research and design: constant innovation, research and development of new technologies and materials, excellent mechanical processing and distinctive design are the focal points of the company philosophy. Thanks to the controlled and constant use of technology, Benelli produces weapons that combine excellent ballistic performance and superb functional qualities. Therefore, the industrial structure and the advanced technology represent essential points of a typical Industry 4.0 framework, which allows to obtain high production standards of every single component to be assembled in the final product.

Among all the various manufacturing processes of the company, the QC phase is a fundamental aspect in the production of a rifle as the finished product must guarantee high performances both at mechanical and aesthetic level. Since the production of the wooden parts is made by external suppliers, the company carries out a quality spot-check in order to prevent that non-conforming items enter to the subsequent stages of assembly, causing slowdowns in the production cycle, and to avoid the selling of a product that does not meet the expectations of the customer. In particular, a very important component of the rifle is the stock, which is assembled in the backside of the weapon, with the task of ensuring comfort and stability during the use of the rifle. As regards the wooden stocks, the analysis on this fundamental part focuses on its aesthetics and surface manufacturing, defining whether it complies with the quality requirements.

3 Problem Statement

The task of QC of wooden stocks is related to assign a certain grade to each item according to the aesthetic point of view. Wooden stocks are characterized by their uniqueness: each one is different from all the others. This difference between all of them leads to a classification problem of the samples according to the aesthetic aspect. It is possible to pass from woods with minimum variations of colour to others with remarkable contrasts between light and dark veins that fit together in a twisted and variegated way. Grade increases as the grain in wood increases, and therefore the item will have a higher value from an aesthetic point of view and, consequently, also from an economic one. Generally, the commercial classification of wooden stocks is defined in five major categories ranging from grade 1 up to 5, where grade 1 indicates almost veinless wood and grade 5 a very twisted and variegated grain pattern. Each different type of rifle model manufactured by the company is equipped with a stock belonging to a specific grade class and this coupling is at the total discretion of the company according to its market decisions.

4 Description of Previous Solutions

Today the aesthetic QC of wooden stocks is only based on the evaluation of the human eye. At first, the operator has to verify the conformity and integrity of the material, then decide whether the item has the right characteristics to be part of the grade class given by its manufacturer. If there is no agreement, the stock can be sent back or reclassified in another grade. As there is no the support of an objective method, the entire process is solely entrusted to the ability and experience of the technical staff, and this implies several main limitations and drawbacks. First of all, the results are affected by a high subjectivity that could give different responses depending on the time when the classification takes place and by which operator. Like all aesthetic evaluations, the personal judgement leads to intrinsic variations in results and therefore alterations are routine, especially on high quality woods. Moreover, a training period is necessary for each operator to acquire the required expertise, and this implies a significant investment in time and resources. Another issue to consider is that the number of samples to be checked depends on the lot size, but the QC is a time-consuming task: for this reason, the operator inspection consists in examining only a minimum part of the stocks delivered by suppliers, following that many pieces are considered right without being evaluated. Although our recent work (Rosati et al., 2021) in this domain proposed a DL approach for solving the QC task, there is still no decision support system which allows the automatization of QC analysis.

5 Material and Method

5.1 Dataset

The dataset was collected by considering both left and right side images belonging to different shotguns, for a total of 1095 images with a size of 1000 × 500 pixels (Rosati et al., 2021). The stocks were classified into 5 different classes according to the company's demand, as reported in Fig. 1 it shows an example of stock for each of the 5 classes. The images were acquired using a dedicated acquisition bench and annotation software.

Acquisition bench

The bench is composed of an industrial lamp and a high-definition RGB camera installed at the top of a photographic box, as shown in Fig. 2. The box shields from external lights: this configuration guarantees the acquisition of images at high

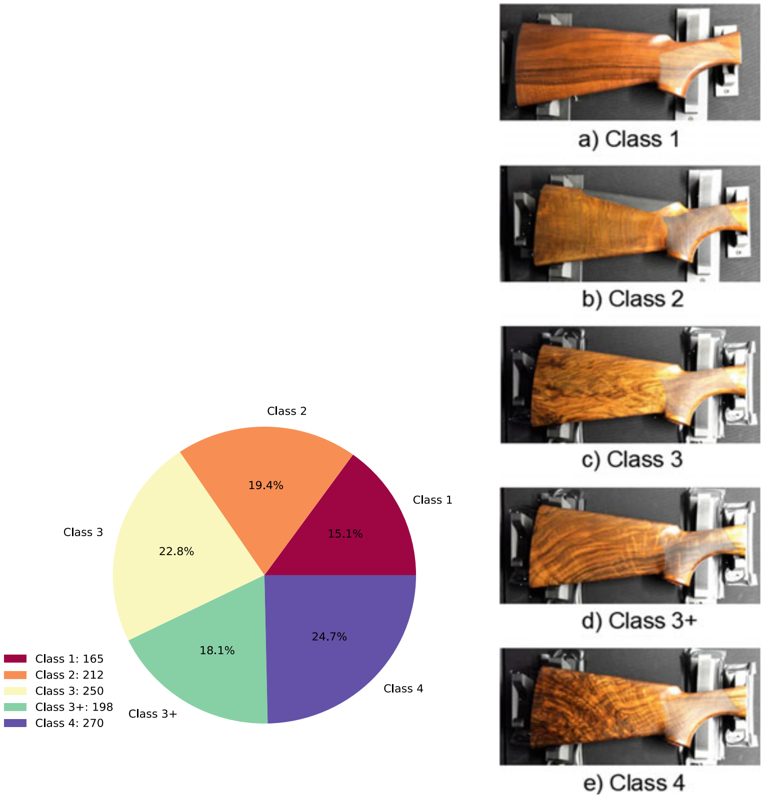


Fig. 1 (Left side) The distribution of the collected dataset among the classes; (right side) Example of different stocks for each of the 5 classes

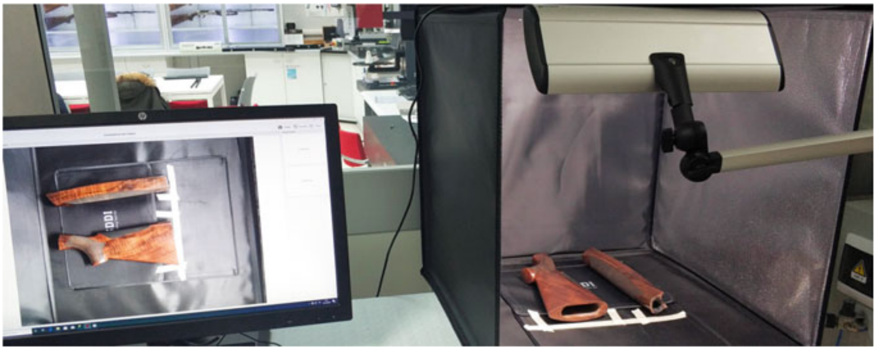


Fig. 2 The acquisition bench installed in Benelli company

resolution with a uniform brightness, avoiding the reflection produced by curvature and polishing of stocks. Once the final configuration had been chosen, the internal staff of Benelli prepared the templates to facilitate the insertion of the items inside the box and to guarantee the homogeneity of the acquisitions.

Annotation Software

A custom software product has been developed to allow operators to acquire images, to note the quality class of the item and to store all in a dedicated database. The user must indicate for both sides of each stock-rod pair: (a) the article typology; (b) the overall quality class between stock and rod; (c) the quality class of stock and rod individually. The interface of the annotation software is shown in Fig. 3. In a second phase, in the same software the predictive models have been integrated so that, when the operator acquires the image, the classification obtained by the models is returned.

5.2 Deep Learning Model

VGG16 architecture

For the classification task of the stocks images, the VGG16 network has been employed, which is a well-known state-of-the-art Convolutional Neural Network (CNN) model. VGG16 network has also been successfully used in our previous work on this domain (Rosati et al., 2021). VGG16 architecture (Simonyan & Zisserman, 1409) is characterized by its relative simplicity, using only 13 convolutional layers that extract image features. Each convolutional block has filters with a 3×3 pixels receptive field and is followed by a ReLU activation function. Reducing volume size is handled by max-pooling. Finally, three Fully-Connected (FC) layers follow the stack of convolutional ones (for a total of 16 weight layers): the first two have 4096 nodes each, the final layer is a softmax layer which performs output classification. The employed network is represented in Fig. 4.

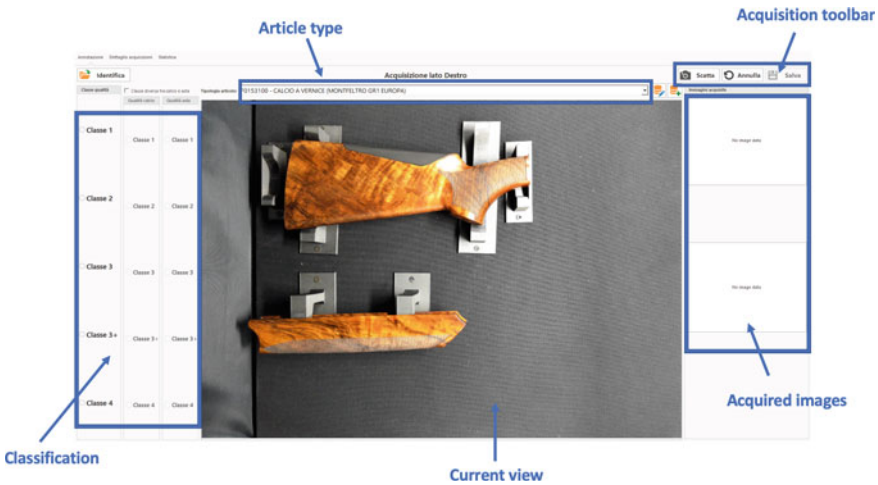


Fig. 3 The custom data annotation software presents several panels: "Classification", where the user selects the overall quality class between stock and rod and the quality class of each item individually; "Article type" for the item typology and the rifle series it belongs to; "Current view" represents the field of view of the camera; "Acquisition toolbar" by which the user can take, cancel or save the image; "Acquired images" which allows a temporary display of the acquisitions

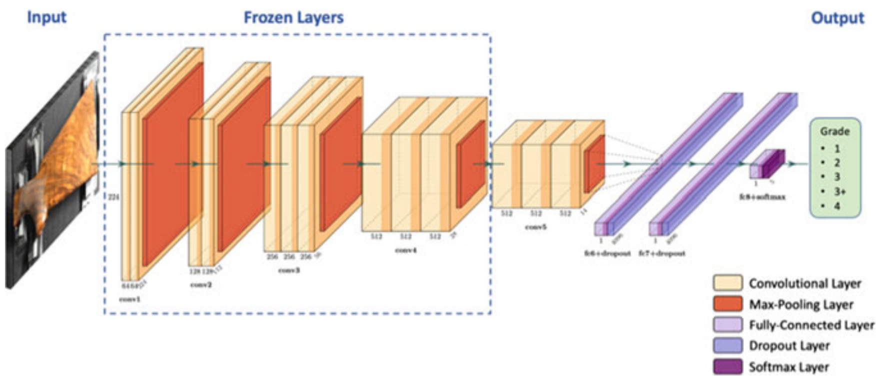


Fig. 4 The pre-trained VGG16 convolutional neural network architecture

Differently from training the model from scratch, we adopted a transfer learning strategy to fine-tune the model on ImageNet pre-trained weights (Deng et al., 2009), while improving the generalization performance. Considering the small size of the dataset, the advantage of this procedure is widely demonstrated in the related literature (Pan & Yang, 2009). For this reason, the first 4 convolutional blocks were frozen and the last fully-connected layer was modified from 1000 (classes in ImageNet) to 5 neurons, which is the dimension of output space for our task. Dropout regularization layers were inserted after the first and second fully-connected layers with a rate of

0.3. Moreover, the images were resized to 224 × 224 pixels in order to match the ImageNet input dimension. The mean value was removed from each image.

Loss function

The standard categorical cross-entropy loss function is defined as follows:

$$\mathcal{L}_c = \sum_{i=1}^{K} t_i logp_i + (1 - t_i)\log(1 - p_i)), \tag{1}$$

where K is the number of quality classes $|Y|$, t_i is the target class vector and p_i is the posterior probability vector. The target class vector is computed by encoding the ground-truth category $y = k$ with $t = (0, \ldots, 0, 1, 0 \ldots, 0)$ where only the element t_k is set to 1. In order to encourage an ordinal structure, we definded a custom ordinal categorical cross-entropy (OCCE) as follows:

$$\mathcal{L}_O = 1 + \Delta \cdot \mathcal{L}_c, \tag{2}$$

where

$$\Delta = argmax\big(t - \hat{t}\big). \tag{3}$$

The main idea behind the loss function defined in Eq. 2 is to penalize the error between the target class vector (t) and the predicted target class vector ($\hat{t}$) according to the ranking (i.e. by penalizing more non-consecutive misclassified samples).

5.3 *Experimental Procedure*

The experimental procedure was performed according to (Rosati et al., 2021). For the training procedure of the CNN model, we adopted the gradient descent (with momentum) as optimizer. We tuned different hyperparameters (i.e. the batch size, the initial learning rate and the momentum in the range $\{32, 64, 128\}$, $\{1 \cdot 10^{-4}, 1 \cdot 10^{-3}, 1 \cdot 10^{-2}\}$, $\{0.8, 0.9\}$ respectively) in a separate validation set using a grid-search approach. The number of epochs was set to 30. The best-selected hyperparameters are respectively 64, $1 \cdot 10^{-3}$ and 0.8 for batch size, initial learning rate and momentum. The dataset was split by a startified hold-out procedure, i.e. using 60% of images as training, 20% as validation and 20% as test. Images belonging to the same shotgun ID were maintained in the same set. This checking was performed to ensure that the algorithm may be able to generalize across different unseen shotgun stocks. Data augmentation was performed on-the-fly on the training set, applying horizontal flip to original images, in order to cope with the small dimension of the dataset. All the experiments were performed using TensorFlow 2.0 and Keras 2.3.1 frameworks on

Intel Core i7-4790 CPU 3.60 GHz with 16 GB of RAM and NVIDIA GeForce GTX 970.

5.4 Performance Evaluation

The metrics used for the performance evaluation of the network are:

- ***Accuracy***

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN},$$

- ***Precision***, also called *positive predictive value*

$$Precision = \frac{TP}{TP + FP},$$

- ***Recall***, also known as *sensitivity*

$$Recall = \frac{TP}{TP + FN},$$

- ***F1_Score***

$$F1_{Score} = 2 * \frac{Precision * Recall}{Precision + Recall},$$

where TP is True Positive, FP is False Positive, TN is True Negative and FN is False Negative.

6 Results

In Sect. 6.1, we reported the performance of the VGG16 model with OCCE loss function for solving the classification task compared to other state-of-the-art networks (i.e. AlexNet (Krizhevsky et al., 2012) and ResNet50 (He et al., 2016)). Afterward, in Sect. 6.2 we described how we have detected and minimized unwanted bias. Finally, in Sect. 6.3 we show the result of the blind test executed by the dataset annotator in order to validate the DL model.

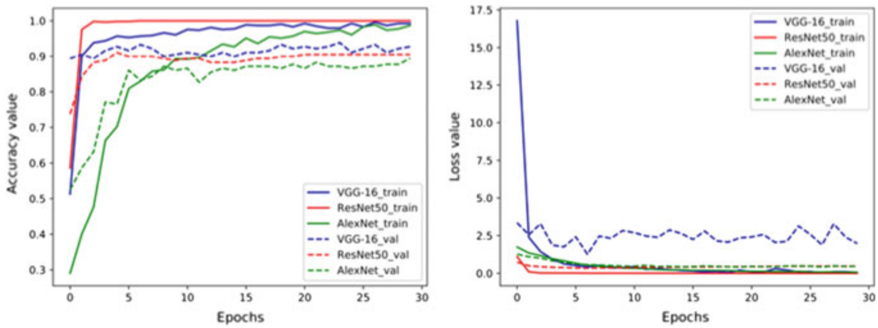


Fig. 5 (Left side) Accuracy curves and (right side) Loss curves across each epoch during training and validation phases for VGG-16 model compared to AlexNet and ResNet50

Table 1 The values of accuracy, precision, recall and F1-Score of VGG16, AlexNet and ResNet50 networks among all the classes

CNN model	Accuracy	Precision	Recall	F1_score
VGG16	0.86	0.85	0.86	0.86
AlexNet	0.83	0.84	0.83	0.84
ResNet50	0.66	0.74	0.68	0.71

6.1 Comparison with Other State-Of-The-Art CNN Models

Figure 5 (left side) shows the training and validation accuracy of VGG-16, ResNet50 and AlexNet across each epoch for solving the multi-class classification task. In all cases, the training time was about 25 s per epoch, with an inference time of 3 s. The validation accuracy of VGG-16 overcomes both that of the ResNet50 and AlexNet. It is worth to mention here that all models seem to be early affected by overfitting. Table 1 shows the classification performance of three models using the OCCE loss on the test set. The VGG-16 overcomes the ResNet50 and AlexNet in terms of Accuracy, Precision, Recall and F1. This is also confirmed by the confusion matrices reported in Fig. 6.

6.2 Bias Detection and Mitigation

Starting from the assumption that each specific rifle series is linked to a specific quality class according to the design defined by the company, we investigated the correlation between rifles series and ground-truth quality classes. The Cramer's V Correlation coefficient is equal to 0.67 ($p < 0.05$). Moreover, considering also that each series has different exclusive characteristics (such as size, shape, colour, polishing, plastic insert), this may lead the model to focus more on the structural

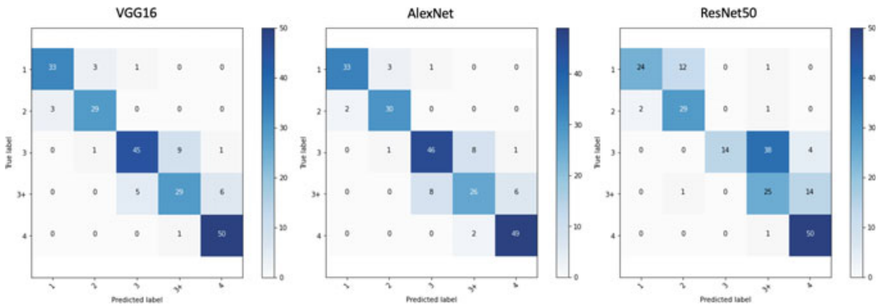


Fig. 6 Confusion matrices of (left) VGG-16, (middle) AlexNet and (right) ResNet50 models

characteristic of the stock than on the wood grain. In fact, the Cramer's V Correlation coefficient between rifles series and the VGG-16 predictions is equal to 0.70 ($p < 0.05$). With the purpose to constrain the DL model to the evaluation only of the wood, we have cropped all the images in order to focus only on the central part of the stock, as reported in Fig. 7.

Figure 8 shows the training and validation accuracy and loss curves of the VGG-16 model trained with the cropped images for solving the multi-class classification task. These curves show less tendency to overfit than the previous model. The predictive performance of the VGG16 is reduced with respect to the previous case, above all for the classes 3 and 3 + that are exchanged since their high similarity as demonstrated by the confusion matrix in Fig. 8 (right side) and in Table 2.

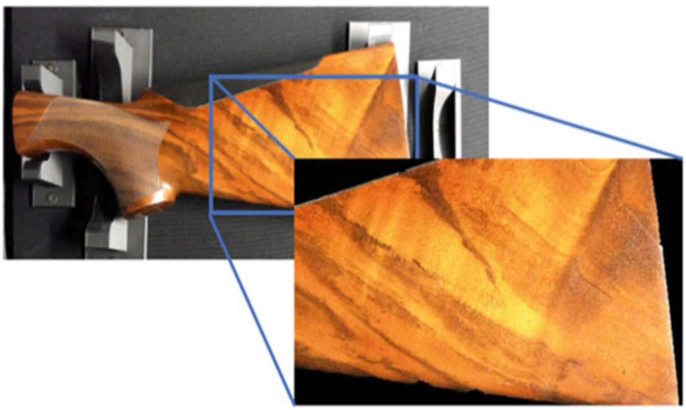


Fig. 7 An example of the cropping performed on the images of the dataset

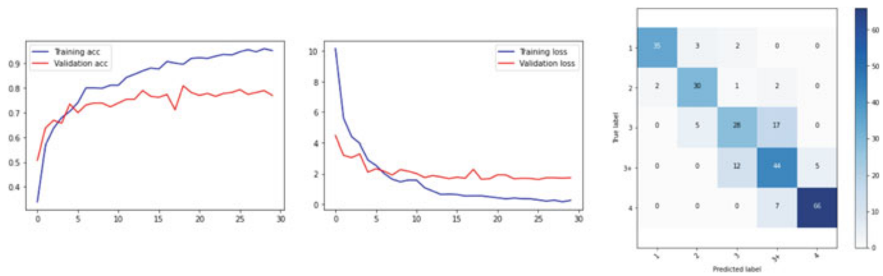


Fig. 8 (Left) Accuracy curves and (middle) Loss curves across each epoch during training and validation phases for VGG-16 model with cropped images. (right) Result of the confusion matrix for the same model

Table 2 The values of per class precision, recall and F1-score of VGG16 network with cropped images. Accuracy value: 0.78

Classes	Precision	Recall	F1_score
1	0.95	0.88	0.91
2	0.79	0.86	0.82
3	0.65	0.56	0.60
3 +	0.63	0.72	0.67
4	0.93	0.90	0.92

6.3 Validation by Human Annotator

After the development of the model, it has been validated with a blind test by the same operator who has annotated the employed dataset. A subset of 18 test images in which the model predictions were different from the ground-truth classes was selected. The operator evaluated the images without knowing the grade he had previously assigned and the prediction of the VGG16 model. The validation results were reported in Table 3. It is interesting to note that only in 5 cases the operator repeated the same classification, proving that the task is very challenging even for an expert human eye. Moreover, in 9 cases out of 18 the new classification corresponds to the same one provided by the model.

7 Discussion and Conclusions

Taking into account the high variability among different wooden stocks, the aesthetic quality classification of rifles stocks based on the analysis of wood grains represents a challenging and relevant step during the overall production chain. The function of quality control (QC) is to build a method that can make objective the result of the inspections carried out, which are still purely dependent on the evaluation of human operator (inter-operator and intra-operator variability). Thus, as stated above, the

Table 3 Result of the blind test. Red rows: cases where validation class is equal to training class. Green rows: cases where validation class is equal to the class predicted by the VGG16 model

ID_stock	Train_class	VGG16_prediction	Validation
1865	3+	3	4
1875	1	2	2
352	2	1	2
242	3+	3	3+
586	3+	3	3+
27	3+	4	4
220	3+	3	3
282	2	1	2
1650	1	2	2
267	1	2	2
644	3	1	3+
201	1	3	3
54	3+	4	4
1887	1	3	2
3	3+	4	4
7	3+	4	4
241	3+	3	4
276	2	1	2

aim is to create an intelligent system for the automation of this aesthetic assessment of wooden stocks using DL techniques, with the purpose of making this control more reliable, fast and standardized. Being trained on examples annotated by experts rather than composed of strict descriptive rules, the proposed DL approach are able, with the necessary training, to generalize across different unseen shotgun stocks. However, the advantage of the proposed approach is not limited to the automatization of the overall QC procedure and the minimization of the annotator's variability. Applying traditional DL methodologies (fine-tuning approach on standard state-of-the-art CNN architectures such as AlexNet, VGG16 and ResNet), we found two main bias problems that afflict the results of the aesthetic classification task:

1. considering that each different type of rifle model is related to specific grade class, we noticed that CNNs focus their attention on typical model features (such as shape, colour, size, and special manufactures) rather than the pattern of wood grains;
2. the classes 3 and 3+ are very similar and difficult to distinguish even for an experienced operator. This implies that the task is tricky and a bias could be present in the ground-truth annotations used to train the network.

This last aspect is also evident in the results of the validation phase. Starting from the fact that the annotation, driven by the decision-making aspect of the human expert annotator, represents the ground-truth, the validation procedure performed by the same annotator with a blind test (see Sect. 6.3) highlights the intrinsic difficulty of the task and, at the same time, the goodness of the proposed approach.

The deep understanding of the overall procedure and the strict collaboration with the company lead to perform an in-depth preprocessing step in order to extract only the most relevant discriminative feature for solving the QC task while minimizing any bias factors. The proposed DL approach is able to carry out quality inspection based on DL techniques, with the ability to condense the experience of the best technicians of the reference sectors. This potentiality allows to scale the DL model in different environments and operating conditions. As a result, the potential impact of the proposed approach could be measured according to the (i) introduction of DL technique in order to automatize the QC procedure in a challenging industrial case scenario and (ii) by also ensuring the immunity of the approach with respect to possible bias factors. Moreover, the integration of the trained network in the QC platform allow to reduce up to 90% the time needed for qualitative analysis carried out manually in this specific field. The great flexibility and invariance to environmental conditions of these techniques will also allow a high level of replicability of the project even between companies with production lines characterized by different processes, minimizing the impact on the phases before and after the QC. We are currently measuring how the proposed DL approach is able to generalize across different company's production chain in order to support the QC process of the technician in a different environment and operating condition.

Acknowledgements This work was supported within the research agreement between Università Politecnica delle Marche, Benelli Armi Spa and Sinergia Consulenze srl.

References

Cadavid, J. P. U., Lamouri, S., Grabot, B., Pellerin, R., & Fortin, A. (2020). Machine learning applied in production planning and control: A state-of-the-art in the era of industry 4.0. *Journal of Intelligent Manufacturing,* 1–28.

Calabrese, M., Cimmino, M., Fiume, F., Manfrin, M., Romeo, L., Ceccacci, S., Paolanti, M., Toscano, G., Ciandrini, G., & Carrotta, A., et al. (2020). Sophia: An event based IOT and machine learning architecture for predictive maintenance in industry 4.0. *Information, 11*(4), 202.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A largescale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Escobar, C. A., & Morales-Menendez, R. (2018). Machine learning techniques for quality control in high conformance manufacturing environment. *Advances in Mechanical Engineering, 10*(2), 1687814018755519.

Francis, J., & Bian, L. (2019). Deep learning for distortion prediction in laserbased additive manufacturing using big data. *Manufacturing Letters, 20*, 10–14.

Fu, G., Sun, P., Zhu, W., Yang, J., Cao, Y., Yang, M. Y., & Cao, Y. (2019). A deep-learning-based approach for fast and robust steel surface defects classification. *Optics and Lasers in Engineering, 121*, 397–405.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Iqbal, R., Maniak, T., Doctor, F., & Karyotis, C. (2019). Fault detection and isolation in industrial processes using deep learning approaches. *IEEE Transactions on Industrial Informatics, 15*(5), 3077–3084.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lee, C. Y., Lin, J. Y., & Chang, R. I. (2018). Improve quality and efficiency of textile process using data-driven machine learning in industry 4.0. *International Journal of Technology and Engineering Studies, 4*(1), 64–76.

Li, Y., Zhao, W., & Pan, J. (2017). Deformable patterned fabric defect detection with fisher criterion-based deep learning. *IEEE Transactions on Automation Science and Engineering, 14*(2), 1256–1264.

Muniategui, A., de la Yedra, A. G., del Barrio, J. A., Masenlle, M., Angulo, X., & Moreno, R. (2019). Mass production quality control of welds based on image processing and deep learning in safety components industry. In C. Cudel, S. Bazeille, & N. Verrier (Eds.), *Fourteenth international conference on quality control by artificial vision* (Vol. 11172, pp. 148–155). International society for optics and photonics.

Ozdemir, R., & Koc, M. (2019). A quality control application on a smart factory prototype using deep learning methods. In *2019 IEEE 14th international conference on computer sciences and information technologies (CSIT)*, vol. 1 (pp. 46–49).

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.

Peres, R. S., Barata, J., Leitao, P., & Garcia, G. (2019). Multistage quality control using machine learning in the automotive industry. *IEEE Access, 7,* 79908–79916.

Romeo, L., Loncarski, J., Paolanti, M., Bocchini, G., Mancini, A., &Frontoni, E. (2020). Machine learning-based design support system for the prediction of heterogeneous machine parameters in industry 4.0. *Expert Systems with Applications, 140*, 112869.

Rosati, R., Romeo, L., Cecchini, G., Tonetto, F., Perugini, L., Ruggeri, L., Viti, P., & Frontoni, E. (2021). Bias from the wild industry 4.0: Are we really classifying the quality or shotgun series?

Sakib, N., & Wuest, T. (2018). Challenges and opportunities of condition-based predictive maintenance: A review. *Procedia CIRP, 78,* 267–272.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Available at: arXiv:1409.1556

Villalba-Diez, J., Schmidt, D., Gevers, R., Ordieres-Meré, J., Buchwitz, M., & Wellbrock, W. (2019). Deep learning for industrial computer vision quality control in the printing industry 4.0. *Sensors, 19*(18), 3987.

Wang, T., Chen, Y., Qiao, M., & Snoussi, H. (2018). A fast and robust convolutional neural network-based defect detection model in product quality control. *The International Journal of Advanced Manufacturing Technology, 94*(9–12), 3465–3471.

ML & AI Application for the Automotive Industry

Antía Fernández-López, Bruno Fernández-Castro, and Daniel García-Coego

Abstract The dawn of Industry 4.0 fosters new opportunities for manufacturing companies in order to improve their operational efficiency and competitiveness through the use of Machine Learning (ML) and Artificial Intelligence (AI) applications. The automotive industry has always been a leading industry in applying new methodologies and R&D results to all the steps of its value chain. Due to the variety and typology of data available through these different steps, new data processing architectures shall be applied in order to empower novel data analytics applications. This chapter presents three use cases related to the application of ML and AI solutions in manufacturing processes from the automotive sector: (I) Real-time quality prediction, (ii) one-off raw material optimization and (iii) real-time industrial robot anomalous behaviour detection. Data analytics allow to identify in advance the generated product quality and speed up the production process and reduce waste, while ensuring the quality of all parts generated for each batch. In this case, this system is working in real-time while the plant is being producing. In other cases, the analysis is done once and there is no need to re-measure the parameters in real time. This is the case of the identification of the lower hydrogen consumption for a minimum product quality accepted by the client. Some industries include robots from different manufacturers. The management becomes complex due to the isolated reporting platforms. A system is presented for capturing data in real-time, analyzing and warning anomalies.

A. Fernández-López (✉) · B. Fernández-Castro · D. García-Coego
Galician Research and Development Center in Advanced Telecommunications (GRADIANT), Carretera do Vilar, 56-58 36214 Vigo, Spain
e-mail: afernandez@gradiant.org

B. Fernández-Castro
e-mail: bfernandez@gradiant.org

D. García-Coego
e-mail: dgarcia@gradiant.org

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8_4

1 Introduction

The manufacturing sector has been immersed in a digital transformation towards the Fourth Industrial revolution for several years. The latest technological advancements allow the whole value chain to be more efficient, but the focus should also be placed on factories and industrial processes, the real protagonists of this change. How they achieve this transformation while remaining competitive and profitable in a more and more demanding market while creating new business models is one of the main questions to be analysed. Also, the new scenario brought in by the COVID-19 pandemic has created new challenges for manufacturing companies, as they need to face disruptions in the supply chain and demand, new problems related to the safety of workers and more and more digitized scenarios that arise due to remote working.

In recent years, a new paradigm that aims to achieve operational excellence from digital transformation has arisen: Industry 4.0. Through the application of Key Enabling Technologies (KETs), the manufacturing industry can improve its productivity and efficiency, obtain a more extensive knowledge from their processes and generate new business models supported by digital tools. Among those KETs, Artificial Intelligence (AI) and Machine Learning (ML) are the ones with higher expectations and potential impact on several areas such as operations management, design, quality or maintenance.

Artificial Intelligence and Machine Learning are considered one of the main driving forces towards realizing the following industrial revolution. In conjunction with Cyber-Physical Systems (CPS), they help to build automated systems that learn on their own and allow to improve decision-making and daily tasks of operators, managers and executives. Predictive maintenance, generative design or the optimization of supply chain operations are only some of the many use cases where the use of AI and ML in the manufacturing sector can be applied in order to foster the evolution and competitiveness of industrial companies.

The first step towards achieving this level of intelligence is monitoring and extracting data from the different assets that are deployed not only in the shop floor but also in other departments of the company such as Human Resources, or through the different steps of the supply chain. Previous systems were automated to work independently from each other, but at the moment the integration of intelligent systems in a networked environment allows them to collaborate and combine data in a way that enables knowledge generation and insight creation in a way that could not be achieved in the past. Thus, once information silos are eliminated, AI and ML are able to create new relationships between data from separate sources, allowing to get information that was hidden among a sea of data or even predicting outcomes based on the interaction between variables where no relation was observed in the past.

Although these technologies have applications in a variety of fields and sectors, the automotive industry has been one of the firsts to apply them in manufacturing.

1.1 The Need for AI and ML in the Automotive Industry

The automotive industry is one of the biggest economic sectors in the world by revenue and also one of the most successful industries of the past century. The worldwide sector turnover can be compared to the size of the sixth-largest economy (Masoumi et al., 2019). Also, in 2019 92 million motor vehicles were manufactured, the first year that the industry faced a decrease in manufacture since the beginning of the 2008 crisis.[1] Its main goal is to design, manufacture and sell motor vehicles (cars, trucks, buses…). However, although this is the final product, manufactured and sold by the main Original Equipment Manufacturers (OEMs) and their after-sales partners, a high variety of suppliers providing all the different parts to build a vehicle support the market. In fact, the majority of the labour force lies on these lower tier suppliers. In the European Union, 13.8 million people are employed by the industry, while only 2.6 million of whom are employed by the OEMs.[2] Thus, the automotive industry is one sector that needs to be constantly pursuing novel and innovative solutions in order to increase its competitiveness and maintain all its base employment.

The automotive industry has always been a leading industry in research and development applied to manufacturing. It has always incorporated the latest advances in daily operations as traditionally it has been a highly demanding sector. Hence, the annual R&D expenditure of the automotive industry is more than 50 billion euros. These expenses are employed in the wide variety of activities performed by the OEMs and their suppliers through all the value chain processes. According to (Hofmann et al., 2017), the value chain of the automotive industry can be described at a high level in the following seven steps, each one with their own AI and ML applications as well as complexities and needs:

A. Vehicle design. This step includes the definition of all the components of the car and both CAD and physical representations of the vehicle. Generative design (Briard et al., 2020), complex physics simulations and optimizations (Oliveira & Fernandes, 2019) are the fields where the application of these novel technologies is being applied in order to improve the design process of automobiles.
B. Procurement. This process includes all the activities related to suppliers and their activities. Classification of suppliers, recommendation of actions for purchasing (Widmer et al., 2019) and demand forecasting (Kim et al., 2019) are some of the examples where AI can contribute to improve the operational efficiency.
C. Logistics. Logistics cover several possibilities in the automotive industry, as logistics with suppliers can be considered but also intralogistics inside the shop floor or OEM facilities. Planning, optimization and simulations are key in order to improve the logistics processes (Markov & Vitliemov, 2020). Also,

[1] Estimated worldwide automobile production from 2000 to 2019, Statista. Online: https://www.statista.com/statistics/262747/worldwide-automobile-production-since-2000/.

[2] Automotive Industry, European Commission. Online: https://ec.europa.eu/growth/sectors/automotive_en.

this process is tightly coupled with production as it directly affects and guides it, so applications regarding stock optimization and forecasting are also one area of application for ML technologies.

D. Production. This process is composed of many subprocesses where the application of AI and ML can greatly improve efficiency and decision-making. Support to the operator through intelligent virtual assistants (Tao et al., 2019), quality control and prediction (Peres et al., 2019), production optimization or anomaly detection (Canizo et al., 2019) are some of the areas of interest for AI and ML in the automotive industry regarding the production process.
E. Marketing. Marketing activities greatly rely on the knowledge of the customer. Thus, learning how the market behaves, customer profiling (Farruh, 2020) to understand its needs and sales prediction (Wachter et al., 2019) are some areas that can be cited where novel algorithms can help to reach more customers or maintain their loyalty.
F. Sales, after-sales and retail. This process is closely related to the marketing step. Thus, several of the previously mentioned activities apply too. Also, in the automotive industry the analysis of customer feedback (Wijnhoven & Plant, 2017) and repairs is quite important as it repercutes in lower levels of the value chain such as design or production.
G. Connected customer. New connectivity solutions and IoT have allowed OEMs to provide new services in their vehicles related to data collection from them, user patterns (Xun et al., 2020) and coordination among vehicles in the road for autonomous driving, among other possibilities.

In addition to traditional problems through the value chain, as described above, the automotive industry will also face new specific challenges that may affect some or every one of these steps. The following list represents a summary of them (Gao et al., 2016).

- Shared mobility and servitization of the automotive market.
- Autonomous driving.
- Electrified vehicles and sustainability.
- Connected services.

1.2 The Role of ICT R&D Centres

Technology Centres are institutions dedicated to the production, dissemination and practical application of scientific and technological knowledge in several fields and industrial sectors (Querejeta et al., 2010). Their main focus is to fill the gap between the long-term research performed in the University and the more academic institutions and the needs of productive companies, which are often of a more immediate nature. Therefore, they should be the link between the most advanced science and the technological needs of the market.

At the same time, increasingly sophisticated technologies are being incorporated into the manufacturing processes, with a growing acceptance of the fact that research and development and innovation activities are key to maintaining the competitiveness of the production base of advanced economies (Berger & MIT Task Force on Production i Innovation Economy, 2013). With the current momentum gained with the 4th industrial revolution and the support provided by digital technologies, many of those innovations are digital in nature, hence the need for entities that support industrial use cases with ICT technologies (Wolfe & Hepburn, 2014).

On top of that gap-filling role, R&D centres also act as key drivers of innovation, especially among sophisticated buyers, through their dissemination activities, technology demonstration events and success story showcasing. These actions serve to foster the awareness of the benefits of advanced technologies in productive sectors, where in many occasions the opportunities for innovation are shadowed by the day-to-day priorities.

This chapter is intended as one of those dissemination documents, trying to showcase some scenarios in which a correct application of ICT R&D in the specific point of a complex manufacturing process was able to vastly improve efficiency or reduce costs, with a fast return on the investment on external services.

Gradiant is the R&D centre that has provided the solutions explained in this chapter. It is a private non-profit foundation that aims to improve the competitiveness of companies by transferring knowledge and technologies in the fields of connectivity, intelligence and security. With more than 100 professionals and 12 applied patents, Gradiant has developed 285 different R&D&i projects, becoming one of the main engines of innovation in Galicia and Spain. During the last years, Gradiant has participated in 22 European projects. In recent years, Gradiant's turnover reached 5 million euros, working with more than 230 companies in 25 countries. Gradiant 's ICT research and innovation activity spreads over a wide variety of market sectors.

2 Data Processing in Industry

Most of the applications and use cases described in the previous section require a thorough procedure to retrieve data and convert it into knowledge that empowers human personnel. In order to achieve this scenario, scalable and flexible data monitoring and processing architectures must be deployed. In lower levels of the automation pyramid, as detailed in the ISA95 standard, the use of Internet of Things (IoT) and Cyber-Physical Systems (CPS) is paramount towards achieving a higher degree of visibility over the shop floor and industrial processes. These novel technologies change the old paradigm of communications between adjacent layers of the automation pyramid and build a new decentralized one, where systems can freely talk between layers. Also, the use of standard protocols such as OPC-UA is also of great importance to enable system interoperability and integration.

In order to enable these flexible architectures, message busses, as depicted in Fig. 1, are used to retrieve data from all the sources (robots, PLCs, machines…) and

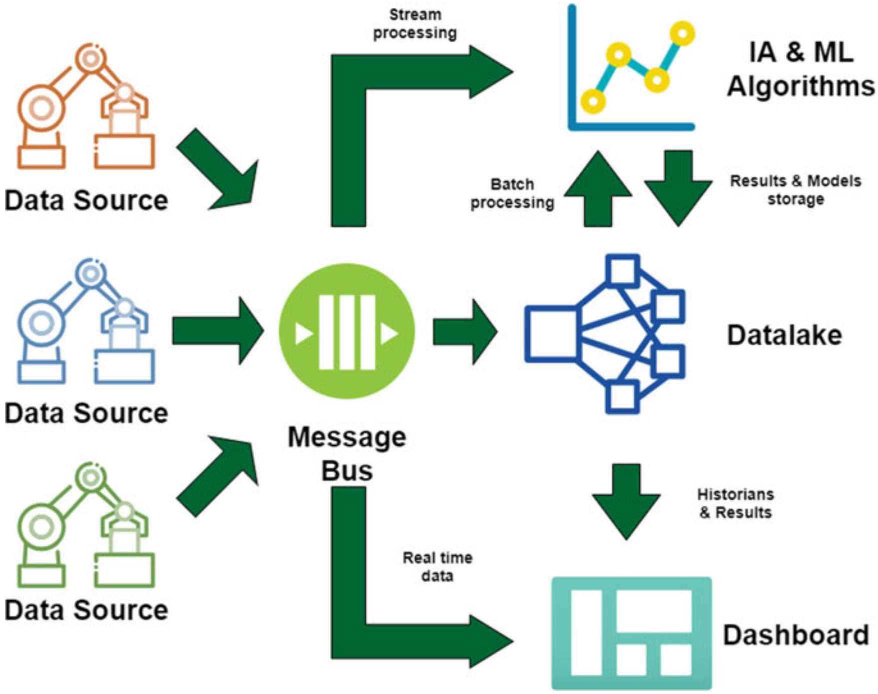


Fig. 1 Data analytics message bus information flow

to route them to the interested actors. The publish-subscribe paradigm is gaining popularity in industrial communications, as it allows to easily connect new data sources, data consuming and decoupling of systems. All the data retrieved is typically stored in a data lake solution, capable of saving all types of data (structured and unstructured, binary, etc.). Analytical services, which can be developed using frameworks such as Keras (https://keras.io/), TensorFlow (https://www.tensorflow.org/) or scikit-learn (https://scikit-learn.org/) among others, may process data streams, directly retrieved from the data sources in real-time, but also in batches that are previously saved in storage systems. Finally, the last step in this chain is to present information to users, typically through the use of dashboards, but also with advanced interfaces such as augmented reality which can also retrieve data in real-time from the data sources or historians and results from the data lake.

3 Automotive Manufacturing Use Cases

3.1 Use Case I: Quality Control

3.1.1 Brief Description of GKN and Its Activities

GKN group is the world's largest manufacturer of constant velocity joints (CV joints) and driveshafts, with a share of global market greater than 40% in such systems. CV joints are design elements for uniform (homokinetic) transmission of torque in both two and four-wheeled vehicles. These elements permit a constant angular velocity on the input, the intermediate and the output shafts.

GKN Driveline Vigo is the largest company within the GKN group in Spain. The factory, with an average workforce of 800 employees, is located in the city of Vigo (Galicia). It has more than 600 machines integrated into 73 flexible production cells, which constitutes a very high intensive plant in discrete manufacturing, producing between 4 and 6% of the whole CV joints sold all over the world (approximately 9.5 million of parts per year).

GKN Driveline Vigo focuses its productive activity on the auxiliary automotive sector, supplying to more than twenty companies (for instance PSA-FCA, Nissan, Ford and Mercedes) in addition to others belonging to the GKN Driveline group.

The factory is also recognized within the GKN group as a global prototype manufacturing centre, in which are developed and validated the operations to be carried out for the intensive manufacture of new designs in the rest of GKN's production plants worldwide.

3.1.2 Problem Statement

CV joints are components manufactured through a complex multistage process, including an induction hardening procedure. This procedure is a key point within the whole production process, with an essential impact concerning the quality of each CV joint manufactured.

The induction hardening treatment process (Cajner et al., 2004) consists of rapid heating of steel parts (CV joints) to a temperature of approximately 950 ºC, followed by an equally rapid cooling procedure (quenching) to decrease its temperature to 25 ºC. The heating occurs as a consequence of a magnetic field created by an inductor (i.e. a copper coil) inside which the joint is located.

The primary objective of these operations is to generate a hardened outer layer whose thickness provides required robustness to each joint. This thickness depends on multiple factors, not only the forging properties of the steel parts, but also other parameters of the process such as the frequency and intensity of the current that circulates through the inductor; the time elapsed in both heating and cooling stages; the geometry of the inductor or the temperature profile, among many others.

The quality control procedure for the induction hardening treatment of CV joints is the result of measuring the accomplished thickness of a randomly selected item drawn from a manufactured batch. This measurement is manually performed by an operator with the support of a hardness tester. The quality of the entire batch will be targeted as defective if the thickness of the selected joint is out of specifications. Nevertheless, this procedure has two important drawbacks: It is a noteworthily time-consuming process and the test joint is completely destroyed in order to carry out its hardness measurements. Taking into consideration the amount of CV joints yearly produced by GKN Driveline Vigo plant, the total waste of raw material is not negligible in terms of production costs. In addition, the procedure does not provide an online quality control of every CV joint after the respective induction hardening treatment, as the quality of the whole manufactured batch is inferred from the measurements of a unique test part.

In this context, machine learning algorithms based on multivariate regression are able to provide valuable outputs in order to overcome the limitations previously described concerning the quality control procedure of CV joints.

3.1.3 The Solution

Multivariate regression is a mathematical approach to model complex dependence relationships between a target variable $Y \in R^n$ and a subset of covariables (or predictors) $X = (X_1, X_2, ...X_p)$ with$X_i \in R^n \forall i \in [1, p]$. The main goal of this type of algorithm is to infer the value of target variables from the values of predictors

$$Y \equiv \widehat{Y} = E[Y/X = (X_1, X_2, \ldots X_p)] = f(X) = f(X_1, X_2, \ldots X_p) \quad (1)$$

The patterns or relationships among covariables, see Fig. 2, represented by function f in Eq. (1), are automatically discovered and modelled by regression algorithms, for example, Lasso (Ranstam & Cook, 2018) or Ridge (McDonald, 2009) for linear dependencies between target and predictors; and Generalized Additive Models (Hastie & Tibshirani, 1990) (GAM) or more complex approaches based on ensembles of decision trees (Bühlmann, 2012), for nonlinear ones.

Supervised methods based on ensembles of decision trees, such as Random Forest (Breiman, 2001; Louppe, 2014), are among the most used techniques to deal with different multivariate regression/classification problems, mainly due to two important factors: An effective procedure preventing the overfitting (Hawkins, 2004) of models, in addition to a high performance obtained with a low computational cost and a reduced set of input data. These methods are based on the premise that the combination of a set of decision trees considerably improves the final results that would be obtained from each one individually.

A decision tree, see Fig. 3, is a structure with two different types of nodes: internal nodes and leaf nodes. Internal nodes are represented by an expression that divides the input data into two disjoint groups according to a split rule

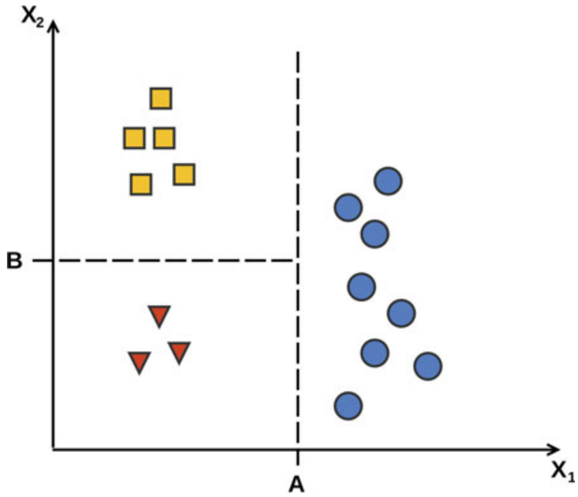


Fig. 2 Two-dimensional dataset with three different patterns (represented by yellow, red and blue samples) and covariables X_1 and X_2

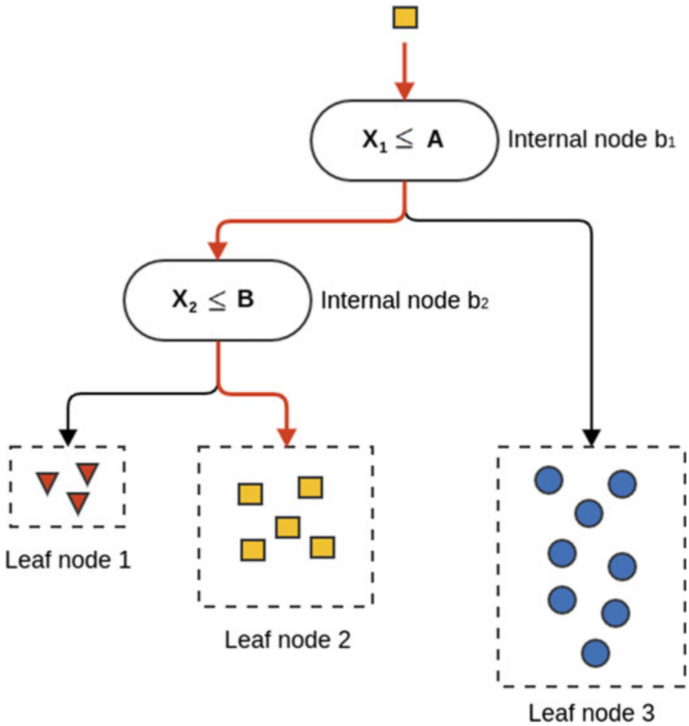


Fig. 3 Example of decision tree learnt during training procedure of a Random Forest regression model. The red line highlights the decision path for a yellow sample

$$X_i \leq \alpha, \forall i \in [1, p] \tag{2}$$

Leaf nodes contain all the samples that follow the decision path verifying the different split rules of the internal nodes from the top to the bottom of the tree. Once a single regression tree is built, the predictions will be obtained as the average of the respective target values related to the samples contained in each leaf node.

Random Forest builds thousands of these trees with different subsets drawn from input data during the training procedure,[3] automatically selecting the split rules (covariables X_i and thresholds α in Eq. (2)) at every internal node (b_j) of each tree that better divides the data, that is, that minimize the prediction error (ϵ) of target variable at the node according to the following expression

$$\epsilon = \frac{1}{n_j} \sum\nolimits_{\forall i \subset b_j} (y_i - \widehat{y}_i)^2 \tag{3}$$

where $\widehat{y_i}$ in Eq. (3) represents the prediction of target sample y_i

Once trained, the prediction of a new sample ($\widehat{Y}_i$) will be calculated as the arithmetic mean of all single predictions obtained from the M trees in the ensemble ($\widehat{Y}_i^k$)

$$\widehat{Y}_i = \frac{1}{M} \sum\nolimits_{k=1}^{M} \widehat{Y}_i^k \tag{4}$$

With the aim of overcoming GKN's problem concerning the quality control procedure for the induction hardening treatment of CV joints, a multivariate regression model based on Random Forest was trained. The model estimates in real-time, and without any waste of raw material, the thickness of each CV joint (the target variable) from the induction hardening treatment process parameters monitored during its fabrication (covariables or predictors): electrical parameters of the inductor such as frequency, current, voltage, apparent power, active and reactive power; and parameters of the heating procedure like the exposure time of every CV joint to the electromagnetic field, in addition to its energy.

The accuracy of the regression model is commonly measured by an evaluation metric called coefficient of determination (Nagelkerke & Others, 1991) or R^2, defined as Eq. (5)

$$R^2 = 1 - \frac{mse(\widehat{y})}{var(y)} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \underline{y}_i)^2} \tag{5}$$

This metric computes the rate of the estimation errors ($mse(\widehat{y})$) with respect to the variance of the target variable ($var(y)$), providing important insights regarding the

---

[3] Random Forest performs bootstrap in training data to reduce the variance of the input samples in order to prevent overfitting.

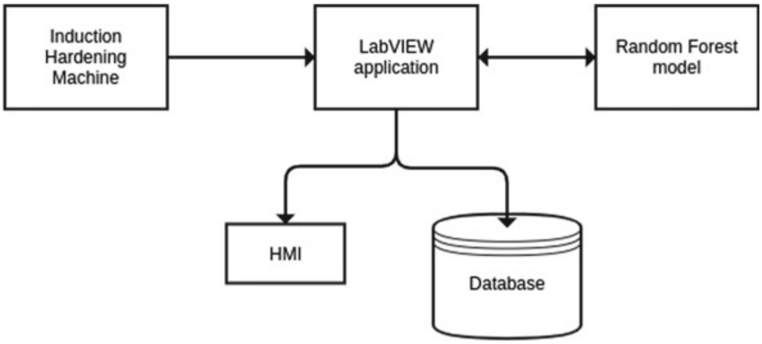


Fig. 4 High-level architecture of the solution developed

proportion of input variance (that is, information related to thickness of the joints) the model is able to explain or predict. The coefficient of determination is generally expressed in percentage, and usually ranges between 0% (the worst of the possible models) and 100% (the best one).

After experiments with different datasets including samples with production parameters and their respective thickness values, the trained model was able to successfully predict around the 92% of the target variance.

The model, coded in a python class with scikit-learn library (https://scikit-learn.org/stable/index.html), was integrated within a LabVIEW (https://www.ni.com/labview) software application, installed and deployed in an embedded system close to the induction hardening machine for edge computing. This application automatically collects the proper parameters monitored during the fabrication of each CV joint and executes the model in order to obtain the estimation of its thickness in real time. The information related to the performance of the process (both production parameters and model's outputs) is presented to the operators through a Human–Machine Interface (HMI) as well as stored in a database for its long-term persistence. The architecture of the solution is depicted in Fig. 4.

3.1.4 Impact of the Solution

The solution implemented at GKN's Driveline Vigo plant to solve the problem statement previously described is a valuable example of how machine learning can satisfactorily contribute and notability improve real-time quality control procedures in high intensive manufacturing processes. The main contributions of this type of algorithm focus not only on online traceability of production quality, but also on a significant reduction of fabrication costs through both the avoiding of traditional destructive tests and the reduction of resources required, in terms of time and equipment.

3.2 Use Case II: Resource Optimization

3.2.1 Brief Description of BorgWarner and Its Activities

Borgwarner's main activities are transforming vehicles to more fuel efficient, cleaner technologies. They have manufacturing and technical facilities in 65 locations in 17 countries, and they employ approximately 27,000 worldwide. Their expertise provides the critical underpinnings of efficient and powerful propulsion. Integrating electronics into the mechanical system is the key to performance, packaging and cost. Borgwarner is a leader in clean, energy-efficient propulsion system solutions. They create combustion, hybrid and EV products for light vehicles, medium and heavy-duty vehicles as well as off-highway applications.

3.2.2 Problem Statement

Certain vehicle components manufactured by BorgWarner are composed of different metal parts which require assembly. These parts are treated with a chemical adherent before putting them in a conveyor belt that moves along a ten-stage furnace at approximately 1000 ºC. Chemical adherent reacts when subjected to the effects of extremely high temperatures, causing the parts to assemble.

In order to prevent the rusting of manufactures, a common flaw in this type of production processes, two mechanisms are employed: parts are slightly sprayed with nitrogen at the entrance and also the exit of the heating process; also, a flow of approximately fifty cubic metres of hydrogen is continuously injected into the furnace with the aim of suppressing the humidity and strictly maintaining the dew point below a threshold of -50 ºC (the dew point is the highest temperature at which the water vapour contained in the air begins to condense producing rusting, among other effects).

Although significantly effective in the prevention of rusting defects, the use of such volumes of hydrogen during manufacturing is inefficient and notably expensive in terms of production costs.

Rusting in metal parts depends on different variables related to production parameters, for instance, the temperature in every stage of the furnace or both humidity and temperature at the shop floor. The offline analysis of manufacturing data collected during production and regarding the condition of the fabrication process provides valuable understanding about the unknown patterns among these factors. These insights are a key contributor for the implementation of a control system to dynamically modulate the flow of hydrogen to the minimum required in order to maintain the dew point under the known threshold, which avoids defects in manufactures.

3.2.3 The Solution

With the aim of reducing the volume of hydrogen required during the heating procedure to avoid rusting, a set of experiments representing different manufacturing scenarios were designed and conducted in a production line at BorgWarners' factory. In each experiment, volume of hydrogen injected in the furnace was specifically modified and monitored together with the rest of production variables: the dew point at shop floor level; the temperature and humidity at shop floor level; the volume of nitrogen sprayed at both entrance and exit of the furnace; the speed of conveyor belt; the temperature in each of the ten stages of the furnace.

Once collected, the time series of data were analysed. The main objective of this analysis is to discover which process variables prominently impact the dew point and, therefore, in the rusting of manufactured parts. To this end, a Random Forest-based multivariate regression model that infers the value of dew point from the rest of variables is trained. From this model, a subset of the most important process variables (i.e. those with the higher predictive power) is obtained through %IncMSE metric.

%IncMSE (Grömping, 2019) is a common unbiased metric to carry out variable importance analysis from Random Forest-based regression models. This metric quantifies how the model predictions decline when a certain covariable changes its value. A higher value of %IncMSE represents a higher variable importance.

For each decision tree $b \in [1 : B]$ in the model, the mean square error of predictions is computed as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{6}$$

where y_i in Eq. (6) represents the values of target variable (dew point) and $\hat{y}_i$ the respective forecastings. Furthermore, each covariable j which has been used to compute each tree $b \in [1 : B]$ is randomly permuted, obtaining a new mean square error of predictions afterwards ($MSE(j)_{permuted}$). The importance of the variable j is therefore defined as

$$\underset{-j}{\delta} = \frac{1}{B}\sum_{b=1}^{B}(\text{MSE}^b - \text{MSE}(j)^b_{\text{permuted}}) = \frac{1}{B}\sum_{b=1}^{B}\delta_{bj} \tag{7}$$

Finally, %IncMSE is obtained as a normalization of Eq. (7)

$$\%\text{IncMSE} = \frac{\underset{-j}{\delta}}{\sigma_{\delta_{bj}}/\sqrt{B}} \tag{8}$$

where $\sigma_{\delta_{bj}}$ is the standard deviation of δ_{bj}.

Table 1 The %IncMSE for the process variables

Process variables	%IncMSE
Humidity at shop floor	40.16
Temperature at shop floor	19.33
Volume of hydrogen	16.34
Furnace temperature at stage 1	12.23
Furnace temperature at stage 10	12.05

The intuition behind this metric in Eq. (8) is as follows: If covariable j does not provide relevant predictive value to the model's output, then the difference between both errors (δ_{-j}), the original and the permuted, should not be substantially larger. On the contrary, if covariable j has predictive value to the response, then the difference between errors should be significantly important.

%IncMSE metric is computed for every dataset obtained from the experiments conducted. The analysis reveals noteworthy understandings related to the process variables. Both the humidity and temperature measured at the shop floor are the most important factors in relation to the dew point, followed by the volume of hydrogen injected as shown in Table 1.

From these insights, a PID (Johnson & Moradi, 2005) controller was installed at BorgWarner's factory, in order to dynamically control the volume of hydrogen injected in the furnace during the manufacturing to prevent the appearance of defects due to rusting of parts. The controller regulates the minimum amount of hydrogen to keep the dew point at approximately -50 °C, also considering the values of humidity and temperature at shop floor, in addition to furnace temperatures.

3.2.4 Impact of the Solution

The savings concerning the volume of hydrogen employed at BorgWarner's factory after the implantation of the controller as a solution for the problem stated were estimated around 13%, which translates into a reduction of approximately 182.000€ in yearly manufacturing costs.

Machine learning techniques are able to achieve important benefits in terms of efficiency and optimization of resources and equipment, but also reveal as a helpful tool to exploit and transform data into valuable information providing important understandings about complex manufacturing processes and their underlyed patterns. This knowledge is critical for decision-making and while designing and implementing data-driven solutions at different stages of the production chain with the aim of accomplishing the best efficiency, effectiveness and performance.

3.3 Use Case III: Data Analytics for Industrial Robots

3.3.1 Brief Description of Ledisson AIT and Its Activities

Ledisson Automation & IT (AIT) is an SME established in the year 2014. Its main objective is to collaborate with OEMs and TIER 1 providers to cover their needs regarding programming of robots and cobots, PLC programming, turnkey integrations or industrial safety solutions. Since 2014, starting from an initial workforce of 7 employees, the company has grown to around 65 employees, 48 of whom are highly qualified technical personnel. Due to a higher implantation of robotics, their main customer is the automotive industry, although other sectors are also covered by Ledisson AIT solutions.

Although initially focused mainly on tasks related to programming of automation assets such as robots and PLCs, due to the arrival of Industry 4.0 and new needs in their customers, Ledisson AIT has started to venture into new services related to added value functionalities for robot monitoring and data analytics.

3.3.2 Problem Statement

As previously commented, in the past, automation systems were deployed in an independent way where interrelationships between them were not considered apart from the different steps in the production process and general efficiency KPIs. Thus, industrial robots were installed with their own siloed control system that relied on sensors or PLCs to know which task to perform repeatedly. Also, data extraction from these systems was difficult due to the lack of standardization, the use of proprietary technologies and the absence of integration with higher level systems. Besides, a typical scenario found on a shop floor is the coexistence of robots from different brands and generations, making it difficult to extract valuable knowledge from their operations in a unified manner.

Industrial Internet of Robots (IIoR) is a product designed to overcome the heterogeneity of robot monitoring systems as well as data ownership issues. It aims to provide OEMs and TIER 1 providers with a scalable and flexible platform to unify data monitoring from their shop floor robots where they can maintain ownership of the data while also providing new data analytics added value services to aim for zero defects manufacturing and efficiency.

Unsupervised anomaly detection is a common issue in the field of machine learning and artificial intelligence. It is defined as the early detection of rare events in a data set, that is, values that significantly deviate from the normal patterns. In robotics, anomaly detection algorithms are applied in order to automatically find out an abnormal behaviour in the movements or actions performed by robots that may feasibly impact on the manufacturing.

Industrial robots usually perform repetitive tasks within certain time slots or production cycles. During these cycles, data regarding the performance of the robot

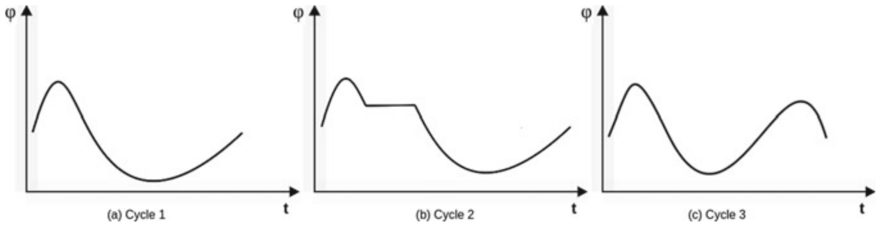


Fig. 5 Example of curves representing variations of angle φ in a robot's axis during three different production cycles: Cycle 1 and 2 are both normal. In Cycle 2 the robot suspends its movement during some seconds; Cycle 3 is abnormal

is collected, for instance angles of each axis or torques, among others. Each cycle is therefore represented by a subset of continuous curves representing the variations of the monitored variables throughout the cycle time. In order to detect abnormal production cycles and subsequent manufacturing defects, the shape of these curves should be analysed and compared. Some examples are shown in Fig. 5.

Nevertheless, in the context of industrial robots, normal curves may present variations that notably alter their statistical properties. A robot slows down or suspends its activity when, for example, an operator approaches the security zone or while waiting for a new item to move forward through the production line in order to be packaged or welded. Even though these variable-length pauses impact on the shape of curves, it does not necessarily mean that a meaningful anomaly occurred. Moreover, the length of the curves is not necessarily equal through all the manufacturing cycles.

All of these issues affect the performance of anomaly detection algorithms for industrial robots in terms of false positive rates, that is, the proportion of normal cycles that are falsely labelled as abnormal.

3.3.3 The Solution

Taking into consideration the problem previously stated, a new unsupervised anomaly detection algorithm for industrial robots was developed. The algorithm is aiming at detecting in real-time abnormal behaviours in robots, through the analysis of curves representing both torques and angles in each axis during a production cycle, which can reveal important malfunctions in equipment or even manufacturing defects.

The algorithm consists of two parts: In the primary stage (training), a different model or baseline representing the normal behaviour of the robot is computed for each variable considered. Each model is trained from a historical dataset with a certain number of curves regarding different manufacturing cycles conducted in the recent past. A model is essentially composed of a reference curve that represents the normal pattern of a robot (for instance, the normal movement of axis number five during a production cycle), in addition to a numeric score or threshold which statistically quantifies how different tend to be the shape of this reference curve from any other normal within the dataset; in the second stage (detection), each new monitored curve is compared with the respective model reference curve in order to detect in real-time significant deviations from normal behaviour. At this stage, the difference between both curves, the new and the reference one, is computed. Providing that the value of this dissimilarity is greater than the model's threshold learnt during the primary stage, the new curve will be labelled as abnormal.

The key point of the algorithm is therefore how both reference curve and threshold are computed during training. To this end, Isolation Forest and Dynamic Time Warping algorithms are employed.

Reference Curve Selection

Isolation forest (Liu et al., 2008) is an unsupervised anomaly detection algorithm based on ensembles of decision trees. Given a dataset of d-dimensional points $X = (X_1, X_2, \dots X_N)$ with $X_i \in R^d \forall i \in [1, N]$, the algorithm attempts to isolate anomaly points rather than profile normal samples. The ensemble is composed of a group of decision trees which are built from a sub-sampling of the original dataset, reducing swamping and masking effects (Ben-Gal, 2005).

As previously mentioned during Random Forest explanation regarding GKN's use case, a decision tree is a structure with two different types of nodes: internal nodes and leaf nodes. Internal nodes are represented by an expression that divides the input data into two disjoint groups according to a split rule defined in Eq. (2).

Nevertheless, in the case of Isolation forest, leaf nodes contain a single sample that follows the decision path verifying the different split rules of the internal nodes from the top to the bottom of the tree. The length of each decision path represents the number of partitions required in the original d-dimensional distribution to isolate each sample from the rest of the points within the dataset. The basis of the algorithm is that a sample with a short number of partitions (i.e. a short length of respective decision path) tends to be more abnormal than those with longer ones. The example can be seen in Fig. 6.

Following this approach, Isolation forest assigns an anomaly score to every single sample x within the dataset X. The score is defined as

$$\rho(x, N) = 2^{-\frac{E[h(x)]}{c(N)}} \tag{9}$$

where $E[h(x)]$ is the average of the decision path length for the sample x ($h(x)$) through all the trees of the ensemble; and $c(N)$ is constant, whose value depends on the number of samples N in the dataset, that statistically normalizes[4] the value of $E[h(x)]$. This score ranges from 0 to 1. The higher the value of the score, the more abnormal the point is.

[4] c(N) estimates the average of the decision path length for every point in a dataset of length N as an unsucessful search in a Binary Search Tree.

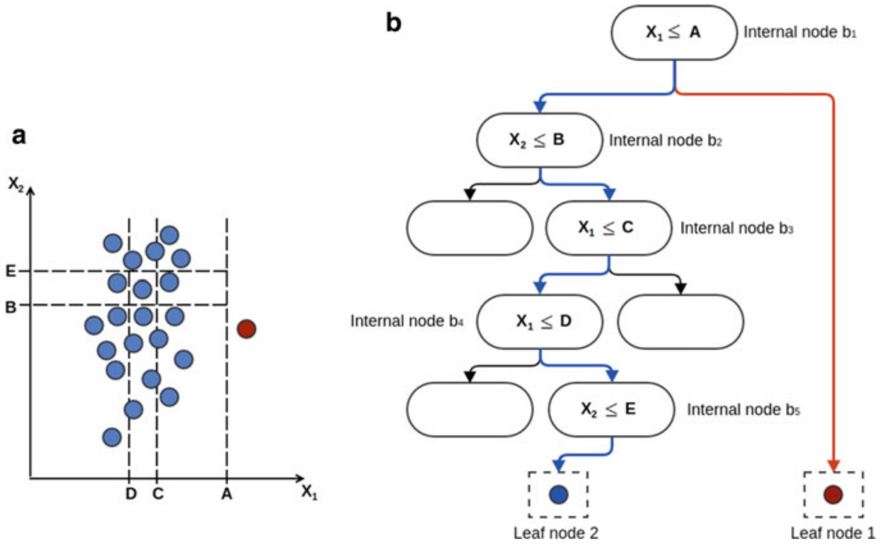


Fig. 6 **a** Two-dimensional dataset with partitions A,B,C,D and E isolating two samples. The sample highlighted in red is an abnormal point. Blue ones are all labelled as normal. **b** Example of decision tree learnt during training stage of an isolation Forest model: The red line highlights the one-length decision path for the abnormal sample; the blue line highlights the five-length decision path for the normal point

An isolation forest-based ensemble of two hundred decision trees is employed in the context of this use case to automatically select the reference curve within the historical dataset as the one with the lowest anomaly score (i.e. the most normal one). To this end, a discretization of each variable-length curve is needed (not all the curves in the training data have to be the same length) as isolation forest requires a fixed dimensional input dataset. This discretization is carried out transforming each curve in a six-dimensional vector with the following discrete variables (Dekking et al., 2005): quartiles of the values along the curve; variance of the curve; range of the curve computed as the difference between its maximum and minimum value; the quadratic mean of the values of the curve.

Threshold Calculation

Once the reference curve is selected, the normality threshold is computed, that is, the numeric score which statistically quantifies how different the shape of this reference curve is in relation to any other normal curve in the training dataset. Common distances between curves, such as Euclidean, could be an acceptable solution. However, as previously mentioned in the problem statement, the shape of two curves representing a robot's performance during two normal manufacturing cycles may be significantly different due to both their variable length and the unexpected

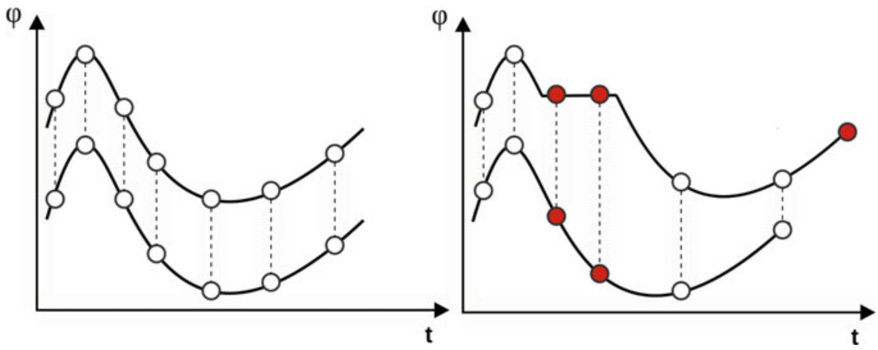


Fig. 7 Left: Euclidean distance between two same-length normal curves with similarity in shapes. **a** Right: Euclidean distance between two different-length normal curves with significant difference in shapes due to pauses in robot's movement. Some points in conflict are highlighted in red

pauses in the robot's performance (see Fig. 7). In such cases, common distances are completely not useful. To overcome this limitation, Dynamic Time Warping (DTW) (Müller, 2007) is employed.

DTW is a time series-based algorithm for measuring similarity between two temporal sequences, $X = (x_1, x_2, \ldots x_P)$ with $p \in N$ and $Y = (y_1, y_2, \ldots y_q)$ with$q \in N$, with variations in length or shape such as robot's performance curves.

The algorithm starts by computing a matrix (D in Eq. (10)) representing the pairwise distance between all X and Y points.

$$D \in R^{pxq} : d_{ij} = \left|\left|x_i - y_j\right|\right|, \forall i \in [1 : p], \forall j \in [1 : q] \tag{10}$$

Once it has built this matrix, DTW automatically finds the correspondences of each $x_i \in X$ to each $y_j \in Y$ that minimize the total distance between both sequences. During this procedure, a point $x_i \in X$ can be associated with multiple points$y_j \in Y$. These correspondences, formally named as alignment path or warping path, are defined as a sequence of points $Z = (z_1, z_2, \ldots z_k)$ with $z_l = (z_{l1}, z_{l2}) \in [1 : p] \times [1 : q]$ for$l \in [1 : k]$, satisfying to the following restrictions:

- Boundary condition: The starting and ending points of the warping path must be the first and the last points of aligned sequences, that is $z_1 = (1, 1)$ and $z_K = (p, q)$.
- Monotonicity: The mapping of the indices from the sequence X to indices from the sequence Y must be monotonically increasing, that is, $z_{11} \leq z_{21} \leq z_{31} \ldots \leq z_{K1}$ and$z_{12} \leq z_{22} \leq z_{32} \ldots \leq z_{K2}$. This condition preserves the time-ordering of points in both time series.
- Step size: Limits the number of shifts in correspondences between points when aligning both sequences.

The DTW distance between curves X and Y, see Fig. 8, is therefore calculated from the warping path as Eq. (11)

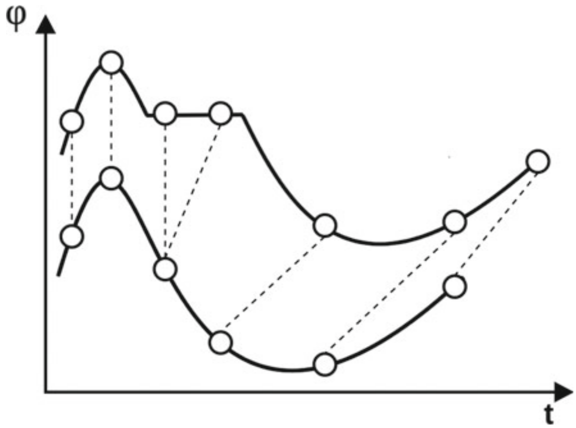


Fig. 8 DTW distance between two different-length normal curves with significant difference in shapes due to pauses in robot's movement

$$D_{DTW}(X, Y) = \sum_{l=1}^{K} \left|\left|x_{z_{l1}} - y_{z_{l2}}\right|\right| \tag{11}$$

From the training dataset composed of M curves regarding the monitored variables of the robot (angles of each axis and/or torques) in different manufacturing cycles conducted in the recent past, the anomaly detection algorithm computes $W = (dtw_1, dtw_2, ...dtw_{M-1})$ with all the DTW distances between the reference curve previously selected and the rest of M-1 curves (dtw_r for$r \in [1, M-1]$) within the dataset. Vector W contains valuable information concerning how historical curves related to normal behaviour of robots differ from the reference one. Finally, the anomaly threshold for the reference curve, which represents improbable deviations from the normal performance, is calculated with a well-known thresholding strategy (Yang et al., 2019) defined as Eq. (12)

$$Threshold \equiv Q_{0.75}(W) + \alpha \cdot IQR(W) \tag{12}$$

where $Q_{0.75}(W)$ is the third quartile of the vector W containing the distances of all robot's curves with respect to the reference one; α is a weighing coefficient with default value to 1.5; and $IQR(W)$ is the interquartile range of W (i.e. the difference between its third and first quartile).

Detection Stage

During the training stage, a different model is computed for each of the robot's variables monitored. Each model is composed of a reference curve that represents the normal pattern of the respective variable, in addition to a threshold providing a maximum level of statistical deviation from the reference shape.

In the detection stage, new unlabelled curves of the performance of the robot are collected. In order to be able to detect in real-time abnormal patterns, DTW distance between the new curve and its respective reference curve is calculated with the aim of measuring their similarity. If the DTW distance exceeds the model threshold, the new curve is labelled as abnormal.

3.3.4 Impact of the Solution

The algorithm developed to overcome the limitations detailed in the problem statement is a serviceable example of a real-time application for unsupervised anomaly detection in industrial robots, specially designed to minimize the false positive rates. The algorithm is able to automatically learn normal patterns in the performance of robots, in addition to detecting rare conditions in manufacturing processes without any support from operators, such as assembling defects or problems during the palletizing of goods.

The early detection of these types of events entails a key improvement in order to optimize the monitoring of fabrication processes and reduce not only production or maintenance costs but also manufacturing flaws. In this context, artificial intelligence and machine learning techniques provide smart-manufacturing solutions to efficiently accomplish these goals.

4 Conclusions

Data analytics applications find application in the entire value chain and every department of industrial companies. Examples of this are vehicle design, procurement, logistics, production, marketing, sales, after-sales and sales and connected customer. Gradiant has created data analytics solutions for many different departments and industries, and this chapter presents three of those use cases related to the production phase in the automotive sector: (i) Real-time quality prediction, (ii) one-off raw material optimization and (iii) real-time robot anomalous behaviour detection. Also, data analytics is presented at a generic level, which covers data sources from different departments and different data analytics systems used for decision-making.

In the case of an intensive and discrete manufacturing plant, the quality control of a single piece for each batch is periodically performed in a destructive process. The presented system made it possible to identify in advance the generated product quality without destroying the sample. This allowed to speed up the production process and to reduce waste, while increasing the confidence in the quality of all parts manufactured in each batch. In order to achieve these results, a data capture system running continuously and in real-time was required, injecting the data into a second analysis system that uses a decision tree to indicate quality.

The identification of the optimum number of resources used in an assembly process is another application. In this case, it was a one-shot data analysis service that offered results which impacted the production line control systems (compared to the previous one that worked continuously, in parallel to the production process). Initially, the relevant variables in the assembly process (quantity of hydrogen, dew point and quality of the product) are detected and in the second step, multiple measurements are made to identify the optimum threshold. After the analysis, the manufacturing process has been adapted with this new number of required resources (hydrogen in the use case) injected into the assembly system, which allows maintaining the quality of the part generated with lower resource waste. The data analytics system is complete, and there is no need to re-measure the parameters in real-time.

Finally, the importance of monitoring the status of various robots involved in a manufacturing plant is explained. Each robot performs a movement defined by the operation that it is executing, and which is captured in real-time by its sensors. This curve can therefore be analysed to detect deviations from an established correct reference, which may indicate robot failures or production defects. The presented system is also able to automatically learn in an unsupervised way as the reference curve changes over time. However, more often than not, a production plant (even a production line) includes many robots from different models and from different manufacturers, all of which usually have non-compatible and independent monitoring platforms. This is why a system to provide a common data access mechanism to different robot families has also been developed and deployed in the described use case.

An important key remark should also be highlighted: Big Data is not an indispensable requirement to implement industrial added value solutions based on machine learning and artificial intelligence. The applications described within this use cases were developed from a relatively low amount of historical data, which did not diminish their usefulness in any way. Normally, the more data available, the more accurate the output of the algorithms are, but this does not necessarily mean that with a moderate amount of data, meaningful improvements cannot be accomplished.

References

Ben-Gal, I. (2005). Outlier detection. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 131–146). Springer.

Berger, S., & MIT Task Force on Production i Innovation Economy. (2013). *Making in America: From innovation to market*. MIT Press.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Briard, T., Segonds, F., & Zamariola, N. (2020). G-DfAM: A methodological proposal of generative design for additive manufacturing in the automotive industry. *International Journal on Interactive Design and Manufacturing, 14*(3), 875–886.

Bühlmann, P. (2012). Bagging, boosting and ensemble methods. In J. E. Gentle, W. K. Härdle, & Y. Mori (Eds.), *Handbook of computational statistics: Concepts and methods* (pp. 985–1022). Springer.

Cajner, F., Smoljan, B., & Landek, D. (2004). Computer simulation of induction hardening. *Journal of Materials Processing Technology, 157–158*, 55–60.

Canizo, M., Triguero, I., Conde, A., & Onieva, E. (2019). Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study. *Neurocomputing, 363*, 246–260.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). *A modern introduction to probability and statistics: Understanding why and how*. Springer Science & Business Media.

Farruh, K. (2020). Consumer life cycle and profiling: A data mining perspective. *In Consumer Behavior and Marketing*. IntechOpen

Gao, P., Kaas, H. W., Mohr, D., & Wee, D. (2016). *Automotive revolution: perspective towards 2030: How the convergence of disruptive technology-driven trends could transform the auto industry*. McKinsey& Company. Available at:http://www.Mckinsey.Com/industries/high-Tech/ourinsights/disruptive-Trends-That-Will-Transform-the-Auto-Industry

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician, 63*(4), 308–319.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. CRC Press.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences, 44(1),* 1–12.

Hofmann, M., Neukart, F., & Bäck, T. (2017). Artificial Intelligence and Data Science in the Automotive Industry. Available at: http://arxiv.org/abs/1709.01989

Johnson, M. A., & Moradi, M. H. (2005). *PID control*. Springer-Verlag London Limited.

Kim, D., Woo, J., Shin, J., Lee, J., & Kim, Y. (2019). Can search engine data improve accuracy of demand forecasting for new products? Evidence from automotive market. *Industrial Management & Data Systems, 119*(5), 1089–1103.

Liu, F. T., Ting, K. M., & Zhou, Z. (2008). Isolation forest. In *2008 eighth IEEE international conference on data mining* (pp. 413–422).

Louppe, G. (2014). Understanding random forests: From theory to practice. Available at: http://arxiv.org/abs/1407.7502

Markov, K., & Vitliemov, P. (2020). Logistics 4.0 and supply chain 4.0 in the automotive industry. In *IOP conference series: Materials Science and Engineering, 878,* 12047.

Masoumi, S. M., Kazemi, N., & Abdul-Rashid, S. H. (2019). Sustainable supply chain management in the automotive industry: A process-oriented review. *Sustainability: Science Practice and Policy*, *11*(14), 3945.

McDonald, G. C. (2009). Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics 1,* 93–100.

Müller, M. (Ed.) (2007). Dynamic time warping. In *Information retrieval for music and motion* (pp. 69–84). Springer Berlin Heidelberg.

Nagelkerke, N. J. D., et al. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

Oliveira, M. C., & Fernandes, J. V. (2019). Modelling and simulation of sheet metal forming processes. *Metals, 9*(12), 1356.

Peres, R. S., Barata, J., Leitao, P., & Garcia, G. (2019). Multistage quality control using machine learning in the automotive industry. *IEEE Access, 7*, 79908–79916.

Querejeta, M. J. A., Wilson, J., Gil, A. M., Varela, G., & Uranga, J. A. (2010). *The Role of Technology Centres in the Innovation Process., 3*(1), 103–108.

Ranstam, J., & Cook, J. (2018). *LASSO regression*. Available at: https://dl.uswr.ac.ir/bitstream/Hannan/45211/1/2018%20BJS%20Volume%20105%20Issue%2011%20October%20%2819%29.pdf

Tao, W., Lai, Z.-H., Leu, M. C., Yin, Z., & Qin, R. (2019). A self-aware and active-guiding training & assistant system for worker-centered intelligent manufacturing. *Manufacturing Letters, 21*, 45–49.

Wachter, P., Widmer, T., & Klein, A. (2019). Predicting automotive sales using pre-purchase online search data. *Federated Conference on Computer Science and Information Systems (FedCSIS), 2019*, 569–577.

Widmer, T., Klein, A., Wachter, P., & Meyl, S. (2019). Predicting material requirements in the automotive industry using data mining. *Business Information Systems*, 147–161.

Wijnhoven, F., & Plant, O. (2017). Sentiment analysis and google trends data for predicting car sales. *ICIS*. Available at: https://ris.utwente.nl/ws/files/20625331/Sentiment_Analysis_and_Google_Trends_Data_for_Predicting_Car_Sale.pdf

Wolfe, D. A., & Hepburn, N. (2014). *Technology and Innovation Centres: Lessons from Germany, the UK and the USA*. Available at: https://munkschool.utoronto.ca/ipl/files/2015/01/Technology-and-Innovation-Centres-Haltech-Report-2014_1.1.pdf

Xun, Y., Liu, J., Kato, N., Fang, Y., & Zhang, Y. (2020). Automobile driver fingerprinting: A new machine learning based authentication scheme. *IEEE Transactions on Industrial Informatics, 16*(2), 1417–1426.

Yang, J., Rahardja, S., & Fränti, P. (2019). Outlier detection: how to threshold outlier scores? In *Proceedings of the international conference on artificial intelligence, information processing and cloud computing* (pp. 1–6).

Applications and Business Impact of Artificial Intelligence in the Industrial Production of Food and Beverages

David Martínez-Simarro and Juan-Pablo Lázaro-Ramos

Abstract The food and beverages industry is required to adapt to an increasingly complex and constantly changing environment and market. Digital technologies, in general, and artificial intelligence, in particular, can play a key role in helping firms in this industry face this challenge and in boosting the quality and safety of their products. This article identifies the main problems arising from the capillarization of digital technologies in society as well as some solutions and methods being used today to guarantee product quality and safety. In view of the limitations of these methods and solutions and to face the changing paradigm for quality and safety in the production of food and beverages, this article describes different experiences in the application of artificial intelligence techniques in various fields, and their impact in business terms for the industry. Finally, the article mentions the main challenges in the application of these technologies in order to make progress beyond the state of the art.

1 Introduction

Today the agrifood sector faces numerous challenges that make it necessary to change many practices and processes in the production and sale of its products, among others: growing market pressure resulting from globalization and increased competition; the need for increasingly differentiated and segmented production; complex requirements for quality assurance; legislative changes requiring more intense processes for quality control and improvement; food safety and traceability requirements throughout the production chain; the need for more reliable and flexible product provision; the need for a better understanding of consumer behavior and trends; and the adoption of sustainable practices.

One of the above-mentioned challenges for the industry stems precisely from progress made in digital technologies, and their extension and capillarization.

D. Martínez-Simarro · J.-P. Lázaro-Ramos (✉)
AINIA Technological Centre, Valencia, Spain
e-mail: jplazaro@ainia.es

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8_5

While such technologies have become cheaper and knowledge is now available on how to develop and adopt them, there has been a radical change in people's eating habits. The market has changed, and new consumers are much more demanding and aware. They have more information to help them decide and are very concerned about what they eat and their way of life. Their needs are increasingly personalized, and they interact with other consumers via various channels. This has led to completely new approach strategies for the organizations that aim to satisfy such new consumers.

Nevertheless, digital technologies have not only changed the way in which consumers reach decisions. They have also changed the structure of the competitive fabric. Such technologies are breaking the rules regarding who has access to interfaces with the market and with market operators, since the digital environment enables producers and consumers to have a direct relationship.

Such changes are, therefore, leading to a transition towards a new model of competition, oriented to the "servitization" or "product as a service". Consumers are increasingly consuming 'services' rather than products, that is, they expect the food products they purchase to add to their quality of life in hedonistic or functional terms. All this is also shaping a new model of collaboration in the value chain and a new productive model. Placing a product on a supermarket shelf is no longer up to a single organization but is determined by a set of them working in, not only on a chain basis but also on a network basis and in a coordinated and synchronized manner to respond fast and effectively to constant changes in preferences and trends. Innovation (in both process and product) has to be increasingly fast and constant. More flexible production mechanisms are needed to orient production not to bulk volumes but to small, even sometimes customized amounts. This requires setting up a production model that is conceived more as ingredient 'assembly' than as a traditional mass production. This is made possible by strong connectivity among the elements involved in production: people, machinery, information systems, and processes that must work in a coordinated way, generating a large amount of information about the production process so that fast decisions can be taken.

This change of paradigm, which is much more complex and dynamic than ever, amounts to a huge challenge for efficiently guaranteeing product quality and safety. Added to this, other aspects such as wide product heterogeneity and the specific nature of perishable products, are also highly relevant. These factors create a further challenge for managing product quality and safety, a problem that is less relevant in other sectors in which production is more homogeneous in terms of product variety.

2 Identification of Sector

The European agrifood industry is made up of a total of 291.000 firms (data for 2020) (FoodDrinkEurope, 2020) and employs 4.82 million workers, which amounts to 15% of all industrial employment. It is the leading industrial sector in terms of volume, added value, and employment. According to figures for 2020 (FoodDrinkEurope), its sales were valued at 1.2 billion euros. The sector is characterized by the great variety of its activities and, consequently, of structures and production methods, as well as great fragmentation, with small and medium enterprises predominating. The European industrial sector, of which the food industry takes part, includes about 2 million firms employing approximately 35 million people and representing 20% of employment in the EU. In spite of these large figures, according to the European Parliament (2015), the European economy has lost one third of its industrial base over the last 40 years and has lost added value. This fact is key for explaining any action in the sector as well as the priorities of firms working in it.

The agrifood industry is heterogeneous and includes firms with a wide range of products of different types although some sectors contribute more than others, such as the meat sector, which represents about 20% of the industry, beverages with about 15%, and dairy about 14%.

The industry also has certain specific characteristics that set it apart from others:

- Its products have to be produced and distributed in safe conditions without pathogenic microorganisms (Mortimore, 2000). This is especially the case for fresh and refrigerated products because poor management of such aspects may result in a public health problem for consumers.
- Its products are perishable (Holley & Patel, 2005), to a greater or lesser extent depending on their specific characteristics and the processes they undergo. It is therefore necessary to monitor their shelf-life both when it is short, as with fresh products, or when it is long as with products that are cured or subject to thermal processes to eliminate micro-organisms or for the purpose of sterilizing their packaging, etc.
- Regulation is very strict in terms of product classification in categories, consumer information and labeling, traceability and product safety control, with different levels of stringency in different markets and with supervisory bodies such as the FDA (Food & Drug Administration, 2021) in the United States or EFSA in Europe, as well as those of each EU member country (European Union, 2021).
- The food production industry is a processing industry. It entails a chain of processes for transformation, addition of ingredients, and processing of a set of raw materials, sometimes also involving microorganisms for biological processes (such as fermentation) or chemical reactions and physical processes arising during production or during the shelf-life of the product until it is consumed.
- Consumers exert a great influence on patterns of consumption and on the demand for certain products and have high visibility on social networks. This forces firms to undertake constant innovation in types of product, processes, formats, and distribution channels, among other areas.

In addition to all the above, food is essential for human life. It is directly related to human health in terms of both availability and its direct effect on people's health when ingested. Moreover, it has to be made available for purchase, whether fresh or processed and whether it has a long or short shelf-life. This requires extensive infrastructure allowing for distribution with great capillarity to all the places where consumers can acquire the products. Food is closely related to peoples' culture and identity and, in some economically more developed societies, it may also take on great importance in a hedonistic lifestyle.

These are the main characteristics of the industrial food and beverage sector. They must be taken into account in the design of any technology if it is to be effective in the sector. Moreover, each sub-sector, such as meat or dairy, may have different needs and product types that set it apart, not to mention the many agents and factors involved in the production, distribution, sale, and consumption of a food product.

3 Problem Statement

The physical, chemical, and microbiological (even organoleptic) characteristics of each type of product (meat, dairy, etc.) vary depending on the food type and change during processing throughout the chain of production. A food is a dynamic system of behavior that varies over time, and there are a multitude of causes for such variations (seasonal/climatic, geographical, species, breed, variety, sex, handling, feed, etc.). Shelf-life or expiration date depends on many factors. Standardizing conditions for production and distribution aims to reduce such variability and to offer products with uniform quality. However, there is an intrinsic variation between batches produced in the same conditions. Processed foods have a diverse final composition which depends on the combination of raw materials, amounting to a complex matrix. Each constituent element has its own behavior and properties and requires specific analytical techniques. Moreover, if quality characteristics are to be analyzed and managed, they must be translated into mathematical language; a specific characteristic (such as weight, color, tenderness) must be expressed numerically so that a population/batch/sample can be described in statistical terms (distribution, mean values, and deviations). From a mathematical point of view, product quality can be said to remain constant as long as the statistical distribution of its quality characteristics does not change.

Certain sensory attributes are subjective and are therefore assessed by tasting panels. Special care must be taken in the selection and training of such panels, the drafting of tasting sheets, the methodology used, and the statistical analysis of the information obtained. Some quality attributes ("experience attributes") can be assessed by the consumer only on purchase or consumption, while others require analysis by specialists rather than consumers ("belief attributes"). To resolve this problem, the latter are presented to the consumer by means of information printed on the label with, for example, quality labels that certify that the producer complies with certain quality standards.

Therefore, the sector has to develop new strategies in order to guarantee safe, quality products in an efficient and sustainable way at the point of sale, in an increasingly complex environment. It is in this context that new advances in artificial intelligence have an essential role to play.

4 Description of Previous Solutions

Artificial intelligence (AI), a discipline that solves problems using a variety of computational techniques, is becoming hugely relevant. Over the last decade, according to the experience of the authors, most of the advances in AI in the food sector have focused on developing models and simulating problems in the behavior of food matrices or of products from the point of view of food safety, always in a laboratory environment and under very controlled conditions. Authors are now observing an explosion in technological proposals and even new applications, in addition to those already mentioned, such as market research and consumption, and analysis of texts in digital publications.

However, today most AI technologies in the agrifood sector are being used in the following 3 main areas:

- Quality control in food production
- Production of food that is safe for consumption
- Compilation of textual information on the market and the sector using natural language processing techniques.

Other areas of action, such as precision agriculture and business analytics, use specific AI techniques for analyzing datasets from the agricultural production environment and from business indicators (sales, customers, market segmentation, etc.), but these fall outside the scope of this article.

The following section covers some of the current solutions adopted in each of the main areas mentioned above, as well as the limitations that are driving the development of new AI-based solutions.

4.1 Limitations of Current Solutions for Quality Control in Food Production

For the agrifood industry, it is important to define the quality attributes that are most highly valued by consumers, as well as their relative importance and how they are assessed. A food is considered of good quality when it meets consumers' needs, which in Europe are food safety, nutritional quality, and healthiness (in the diet as a whole), desirable organoleptic characteristics, and other attributes (environment-friendliness, sustainable development, transparency, and information).

After defining these, a firm's main concern is to achieve continuous production and supply of a product with constantly improving quality. The food industry has drawn up regulations and systems to control and manage quality, including Hazard Analysis and Critical Control Points (HACCP) and Good Manufacturing Practices (GMP). It has also adapted models initially developed for other industrial sectors (ISO standards). The ISO and HACCP systems are preventive and, since monitoring is reported in a set of documents, audits can be carried out and industries standardized.

The HACCP system essentially manages health risks, while ISO standards cover all aspects of quality. There is already a specific regulation (ISO 22000:2005) that lays down requirements and allows for the operation and maintenance of general food safety management systems in the industry.

The HACCP system is used to identify and assess risks, and to monitor in any production process the key points that might affect the safety of a food. There are also regulations set up by large retail chains applying their own safety and quality parameters, such as Global GAP, British Retail Consortium Standard, UNE 155,000, IFS (International Food Standard), and SQF (Safe Quality Food).

Quality control within the industry may not guarantee the safety of the end product because it might be exposed to external changes. For example, for chilled products, it is essential to maintain the cold chain at all times to preserve microbiological stability. This reinforces the fact that distribution and transport operations are an essential part of the agrifood industry and have an impact on product quality and safety.

Quality control in food production may adopt different approaches because of the wide variety of products and of situations. From the point of view of the processing received by products and the raw materials used to make them, products can be grouped as follows:

– Group I (Fresh produce): This group covers fresh foods in their natural state such as fruit, vegetables, meat, fish, and eggs as well as other foods preserved using traditional methods (salting, dehydrating, etc.). Such foods have not been processed and have not undergone any preservation treatment so are very perishable (short shelf-life) and often need refrigeration to prolong their shelf-life.
– Group II: canned or semi-conserved products such as tuna in (oil, mayonnaise, jams), which are usually subject to thermal treatments to sterilize the product and prolong preservation. Usually presented in sealed containers (cans or glass jars).
– Group III: food preserved by freezing or deep-freezing such as fish, vegetables, shellfish, and meat. Defrosting is necessary before consumption.
– Group IV (pre-prepared convenience food): foods such as vegetables, fresh fruit (e.g. sliced mushrooms, cut fruit that is first selected, washed and vacuum-packed or packed in a controlled atmosphere), for which the cold chain must remain unbroken during distribution.
– Group V (pre-cooked convenience food): ready-to eat foods, mainly dishes made up of many ingredients and subject to complex production processes, including thermal processing and packaging in addition to chilling for preservation after packaging.

Different aspects of food safety are stressed depending on the group in which the food is classified. For example, for some foods the emphasis is on evolution of the product's shelf-life and on microbiological control in production conditions (as in a microbiological fermentation process); for others it is on control of the physical and chemical, and even microbiological evolution of the product quality as perceived by the consumer throughout its shelf-life (as in the control of micro-organisms that might affect quality). For other groups or even for types of product within a given group, it may be much more important to analyze the impact of the productions conditions on the product or to understand the influence of the origin of raw materials, or of the packaging, or many other factors.

Whatever the group, in the food sector there are various software applications for quality control, some of which are specific to the food sector while others are more widely used but can be adapted to the characteristics of different types of product. All of which offer basic functionalities such as:

(a) oversight and reports on Control Points
(b) document management
(c) human resources management
(d) management of suppliers and customers
(e) management of audits and non-conforming products
(f) management of equipment maintenance
(g) product traceability
(h) alerts management.

Where laboratory analysis exists, this is usually limited to the checking of certain balanced scorecards as a statistical summary of the main elements of the system. Some of the solutions are connected with production control systems of the MES/MOM type and can draw up graphs of production parameters and alert when pre-set thresholds are passed.

Many firms use ad-hoc solutions to monitor certain quality indicators and HACCP and for real-time production control. However, these do not usually have advanced analytical capability and do not allow product quality control to be connected with the pre-set production conditions and contexts. The basis for analysis is good data collection, and software should have specialized analytical capability and be available for any situations arising.

Another important matter for the food industry is to know in detail how certain key processes perform. For example, software-based simulation mechanisms may be very useful in processes that analyze in detail a given single process in order to find how it really affects the product (as in an emulsion with a fat such as oil, as is often used in sauces).

In the food sector, process simulation (Pieter Verboven et al., 2020) has not evolved as fast or as much as in other sectors, such as pharmaceuticals or chemicals, and still has many shortcomings. One reason for this is that each firm usually devotes resources to research and development precisely because it seeks a unique, differentiated product. This means that processes are very elaborated and streamlined to give a specific organoleptic response in the product or to focus on physical parameters such as size, texture, resistance to pressure, color, and freshness.

There are software applications for the simulation of industrial processes, which simulate chemical reactions, mass transfer processes such as drying for meat or hydration processes, or thermal processes such as those that eliminate pathogens. There are also solutions of generalist simulation software (widely used to simulate multi-parameter physical processes), open source and that can be integrated with third-party software, or other suites, which are widely used in the industry for a variety of physical phenomena. These can give good simulation results provided that the conditions adapt well to the system of equations to be modeled and to the physical and chemical phenomena involved. On a microbiological level, there are modeling suites software platforms, which are available for the product families that are most prone to the micro-biological development of pathogens or to spoilage. Such models describe very well the growth of micro-organisms in controlled conditions but not in a dynamic environment such as production. So, as the production industry in general over the last 10 years and the food industry in particular more recently have gradually adopted computing capability as a result of the availability of data in digital format (Sartal et al., 2019). These are widely used in very intensive, efficient industries such as the automotive or fast-fashion sector, with well-known techniques such as 6-Sigma and agile (Sartal et al., 2017). In the food industry, however, they are being adopted more slowly.

So there are currently no market solutions for process modeling that can be directly and easily adopted to simulate and explore what happens to product characteristics in the food processing industry, in which process interaction is very dynamic.

4.2 Limitations of Prior Solutions for the Management of Food Safety

Among commercial information systems for food safety management, those with the most complete functionalities aim to support HACCP processes. This is a set of procedures to control certain parameters, especially in production lines so that, if certain limits are exceeded (e.g., a product has not remained long enough in a thermal process), alarms are tripped so that users can make any necessary corrections and avoid subsequent risks for the product.

In all of them, control is from the point of view of HACCP, that is, it is reactive. They can connect with probes, sensors, and the information systems of the production system and generate alerts that figure within the actual HACCP plan. However, there is a clear opportunity for progressing in the advanced treatment of the data and for linking the process variables they monitor with the microbiological state of the product. This can add analytical capability to the service provided by the organization's experts, generating a link between the process and food safety.

When updating their systems, firms often introduce software and hardware (probes) or autonomous systems that can emit alerts in the case of changes or trends in a limited number of key variables that are present in their HACCP plan. In all cases, firms either adopt ad-hoc solutions developed by their own IT teams or use generalist solutions.

However, if the problem requires finding out more about what happens in the control of microbiological risks in order to simulate to a greater degree what might happen in the short term on a specific line or a part of it, then a distinction must be made between two large families of mathematical models for assessing microbiological risks in the agrifood context from the predictive point of view:

– Mechanistic models based on equations, also known as white box models.
– Machine learning (ML) models, that is, "black box" models, based on data compiled from the variables that describe the phenomenon.

This type of distinction also applies to the simulation and profound knowledge of the impact of the process conditions on the above-mentioned quality variables.

Various software solutions using mechanistic models are available, both open and commercial. This type of predictive microbiology model is useful to reveal the behavior of microorganisms (presence, trend of populations over time, etc.) in certain food matrices under static conditions, that is, under controlled temperature, humidity, pH, etc. They provide key information on optimal conditions for the growth and presence of such microorganisms as well as on the factors that eliminate them. However, such models cannot be extrapolated to a production environment in which conditions are dynamic. Also, there is interaction between the various processes taking place in the production plant, including not only the production processes themselves but also hygiene patterns, company culture, etc. Moreover, in most cases, the microorganism to be modeled has to be inoculated in the actual food matrix. This cannot be replicated in a real production context because of the effects it might have on subsequent production.

4.3 Limitations for Collecting and Processing Textual Information on the Market and the Sector

There are many social media analysis services available in national and international contexts that can measure, filter, and process publications and perform brand, reputation, and sentiment monitoring. Many of them include mechanisms to quantify the appearance of key words configured by users of the service. Another possible approach is to use commercial advanced text-mining, which offer very advanced text processing techniques that can identify statistical parameters and perform linguistic sentence grouping. Such tools have powerful mechanisms to quantify and extract statistics about impacts (mentions of the terms of interest) and to draw up metrics such as the return on investment in promotion. They can also extract information on the polarity of opinions (sentiment analysis) over time, thus enabling visualization of trends, especially in concepts such as brands.

There are, however, limitations to the commercial solutions available today because, for example, these tools does not take into account the context or the semantic relations between concepts in the agrifood domain in general, nor possible customization in, for example, sauces, industrial baked products or snacks. This is key for correct interpretation of the information captured and expressed freely in social media, blogs, publications, etc. because of the semantic richness with which opinions are expressed about foods, food culture (hedonism, sensory appreciation, etc.), and the impact on people and on their health.

Moreover, most solutions are mainly oriented toward activities related to marketing and product positioning rather than to the generation of an overview of the status of opinion, threats, and opportunities for innovation in a sector like food. Also desirable for the food sector would be qualified associations relating to consumers' assessments of a product based on its ingredients and other food concepts. Such aspects are important, for example, for anticipating reputational crisis (as with palm oil in processed products) or for reacting in time to a change in formulation without waiting for demands from retail or from consumers themselves, or just for detecting potential threats for a specific product or product category imposed by a group of social media influencers.

Although great progress has been made in the development of deep learning technologies applied to the processing of texts in natural language, there are not yet any specialized commercial solutions for detection like those described above, at least nothing specific for the agrifood sector.

5 Description of Current Solutions Related to Small/Big Data and Machine Learning

Having introduced the current situation of commercial technological solutions for managing food quality, microbiological control, and product innovation by active listening in digital media, the following section describes the impact that AI can have on each of these key processes in the agrifood industry. We also describe some real applications on which the authors have been working in recent years.

5.1 Application of AI in Production Quality Control

As already stated, quality control in food production aims to achieve homogeneity, optimization, and efficiency and to reduce complaints and waste in the industrial stage. Given the general complexity of production processes, in which there are great interactions between different processes, modeling them all using mechanistic methodologies based on equations or what are often differential equation systems is very risky because of the accumulated simplification in the models and because such methods do not take into account the interactions among the processes themselves and their effects on the product.

Black-box models (MIT Technology Review, 2020), also known in the industry as Response Surface Modelling (Myers, 2016), collect data on the production environment that are sufficiently variable and can relate decision variables, that is, those that can be controlled by operators or production managers, with the variables that characterize the finished product. Such models can be used, for example, for prediction (or “prognosis”) of product quality on completion in a specific production configuration. The terms “black box” or “response surface” refer to the intuition that the data processing techniques will generate a type of mathematical model that can model the reality that is described by the data but that cannot be accessed or understood by the experts. Supervised machine learning and deep learning, as a customization of the former, are paradigmatic examples of black box models. Their main advantage is that they can model beyond individual processes and deal with very complex problems such as a dynamic production environment. In industrial prognosis, for example, they can predict the quality of a given product. Machine learning techniques such as convolutional neural networks (CNN) (Yan et al., 2019) (widely used in other fields such as image processing or natural language processing) or recurrent neural networks (RNN) (Xia, 2020) can also be used in predictive equipment maintenance.

For global simulation of a line, there is a wide range of techniques that have been tested to different extents in controlled industrial environments but there is not yet a clear consensus on what functions best in different production contexts. Moreover, there are new families of machine learning techniques that have been widely used in automatic learning in images and in natural language, especially in the field of deep learning, such as convolutional neural networks, the introduction of Transformers (Perez et al. 2019) as layers of a deep neural network, or neural network techniques with LSTM (Long Short Term Memory). The first two of these might be very promising for managing sequential information if transferred to other fields, but this aspect has not yet been sufficiently explored in the bibliography consulted by the authors regarding the production environment, and even less so in food production.

5.1.1 Case Study of AI Applied to Food Quality Control

Description of the problem: In a hamburger bun production environment, the main consideration for quality is the evenness of the surface of each unit regarding color continuity and the actual color when cooked. The main problem is that, sometimes and in a way that cannot be controlled by the operators, some batches come out of the oven with a darker color, even though there have been no significant changes in the process variables.

Solution proposed by AINIA (see Fig. 1): Production data were collected over 1 month of production, taking into account variables such as oven temperature, time in the oven, prior fermentation time, and humidity and temperature during baking together with other data coming from different information systems. Taking these data and the quality variable, that is, the surface color, we implemented one machine learning model using neural networks and another using the XGBoost random forest technique.

Results: After various iterations to refine the hyperparameters of these two techniques, we obtained success rates with the neural network model of about 65%, and the random forest model was able to identify a decision tree but none of the variables

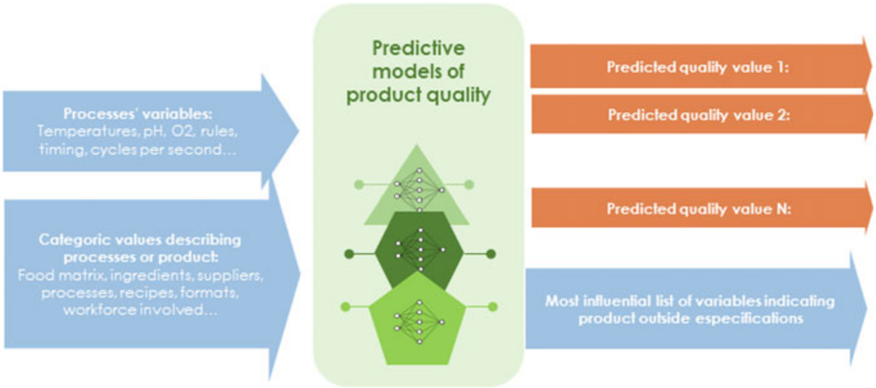


Fig. 1 Schema of the AI solution for predictive models of food product quality

stood out especially. That is, it was not possible to explain the baking phenomenon using the current variables. After conversations with the line operators, we identified the intuition that the oven gas burners might be related to the degree to which the buns were baked. We captured signals on the degree of opening of the 4 burners and on about 300,000 product samples (pans of buns), then again generated and re-trained using the two techniques: neural networks and decision trees. In the first case, we were able to raise the success rate of the prediction of baking to 80%; in the case of the decision tree, we found that the degree of openness of 2 of the burners determined the optimal conditions for baking. After this experiment, the procedure for opening the burners was changed by the company, making it gradual to minimize this effect. This reduced the number of rejected, overcooked buns by almost half.

5.2 Application of AI to the Management of Food Safety Risk

As stated above, industrial food safety risk control is a complex process because many parameters are involved: hygiene protocols, hygiene design of equipment, cleaning protocols for workers and the environment, presence of microorganisms in the environment, temperature and humidity conditions in the plant, origin of raw materials, specific processes to eliminate microorganisms to varying extents, packaging processes, and other variables in the product such as water activity, which may be very relevant for determining risk. Controlling food safety risk in a very dynamic and uncertain environment is very different from the traditional use of predictive microbiology on which current models have been developed, in controlled laboratory conditions. Intuition led us to believe that black box models would be able to integrate all this complexity, which in most cases cannot be linearized. This was a potential technical option although it would probably require a large amount of data samples, depending on the type of microorganisms whose presence is to be modeled.

For black box models, that is, those based on machine learning, several references were found but none for the production environment. They were more an extension of the controlled laboratory conditions of previous models: for example, we found references to prediction of *Listeria monocytogenes* (Gosukonda et al., 2015; Oladunjoye et al., 2017), a pathogen of interest for the industry, with general modeling using neural networks in controlled production conditions (Fernández-Navarro et al., 2010), and risk prediction with machine learning associated with HACCP data in a meat firm (Kaiyi et al., 2007).

Machine learning has also been used in the context of food safety, not to identify production risks but to detect the existence of microorganisms or of chemical spoilage in the actual product. Other applications of machine learning come from the field of natural language processing and aim to identify public health events from text analysis (Geng et al., 2017, 2019) in digital media of different types.

A field on which AINIA is working is the exploration of current deep learning techniques with the aim of building predictive models that can represent with sufficient expression the phenomena that arise throughout production and shelf-life, and their effect on the presence/growth of microorganisms. Among the techniques analyzed and tested, there are some that worked well in contexts such as natural language processing, in which a time component was introduced in the information coming from the samples:

- Long Short Term Memory neural networks (LSTM (Hochreiter & Schmidhuber, 1997)) are a sort of Recurrent Neural Network (RNN (Mikolov et al., 2010)). They are characterized by the fact that it is possible to do feedback among their perceptrons, generating a certain memory in the network itself. They are thus ideal for processing and predicting time series, which are some of the elements that arise in data collection in a production environment. LSTM are being used successfully to predict product demand (Abbasimehr et al., 2020) and prices, but no bibliographical references have been found applied to the prediction of microbiological risk in a production plant environment.
- Back Propagation Neural Networks (BPNN). One study has been found in the bibliography (Deng et al., 2019) using these to predict the presence of coliform bacteria in food, in the framework of a multidimensional problem and with categoric (non-numerical) information such as the representation/presence of certain ingredients and variables such as pH and temperature.
- Convolutional Neural Networks (CNN). In recent years, these have been used in the field of deep learning in image recognition, because they can break down the detection problem into specialized layers that decide about specific characteristics (e.g. contours, fillings, blurring, etc.). In the area of food, they have been used, for example, to classify fruit (Pan et al., 2017) or to identify food safety problems (Jiang et al., 2019).

The above techniques are the most promising in the state of the art of deep learning and, for the current prediction problem; research is focusing on identifying which of them can achieve the best results. These may differ from one firm to another. Although traditional machine learning techniques are not being set aside such as the initial working approach, nor others that might look promising throughout the life of the research project, the results of such techniques (e.g. assessment using ROC curves) will be tested and compared, iterating tests with the processed data and hyperparameters until optimal predictive capability is achieved.

When supervised machine learning or deep learning techniques are used, the development of models is associated with a set of data obtained under certain conditions using a set of techniques and their associated hyperparameters. This means that subsequent use of the model with, for example, predictive purposes, must be done on a system that generates data under the same conditions as those under which the original data were generated. For production, this means that, if we capture information on a specific production line with its unique characteristics, the model will not function if tested on another line, which amounts to a wasted investment on generating the samples in terms of both time and cost. This, precisely, is the context in which Transfer Learning (Pan & Yang, 2010) becomes necessary. The aim is to have a trained network with a prior data set, which can be re-trained with specific data sets. One of the main advantages is that, if we start out with a model trained in the same area, the number of samples to be included in the new 'transferred' model will be smaller than if we were to develop the model from scratch. Classic examples of this type of technique are Word2vec (Tensorflow, 2022) in the field of natural language processing or VGG-19 (Kaggle, 2017) in image processing.

Another type of technique that is potentially effective is Reinforcement Learning (Nguyen et al., 2020) Here the aim is to make an agent learn from actions performed and feedback provided to it on whether the action was correct or not (Sutton & Barto, 2018). Such techniques are being used for learning in gaming. The intuition is that a production environment can be governed like a game, that is, decisions can be taken on consignments, reference units, alarms, process times, etc., and the effect is the potential appearance of microbiological risk in the finished product or during its shelf-life. Some of the algorithms associated with such techniques are Q-learning (Melo 2001) and SARSA (Dilipkumar, 2020) (State-Action-Reward-State-Action). Reinforcement learning techniques can also be used, for example, in problems where there is not a large number of samples, and if the aim is to reinforce or penalize some samples but not others.

5.2.1 Case Study of AI Applied to the Management of Microbiological Risks

Description of the problem: On a production line for Group 4 vegetables (lettuce, cabbage and other cut and prepared salad vegetables), there may be episodes when *listeria monocytogenes* could be present. Current HACCP methods cannot identify the main cause of this presence in the production process. Since the product is fresh, sterilization processes (such as heat treatment) are not possible because they would impair the product. The aim is to identify the current risk (high, medium, or low) that the product might contain the microorganism in the current conditions on the line.

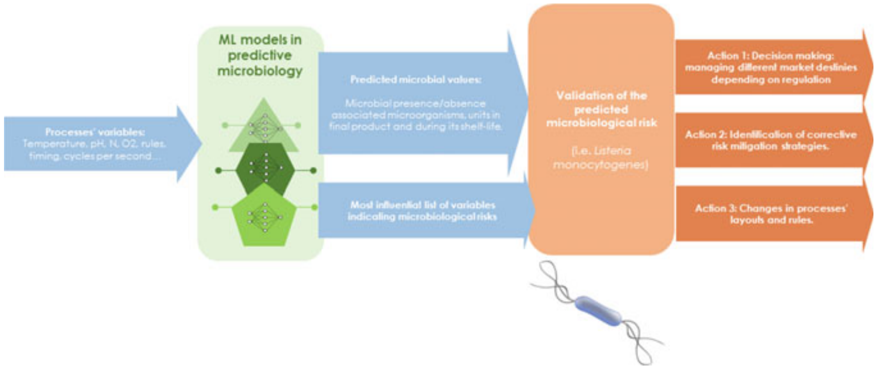


Fig. 2 Schema of ML models in predictive microbiology

Solution proposed by AINIA (see Fig. 2): The proposal was to capture process data linked to the presence of the microorganism and environmental data that might contribute to it: ambient temperature, temperature and pH of washing water, drying process control, percentages of gases during controlled-atmosphere packaging, original supplier of the raw material, and types of raw material. For example, we had the intuition that certain vegetables 'hide' listeria better than others did and, of course, the presence/absence of the target microorganism captured using laboratory techniques, or of other harmless microorganisms that behave in a similar way. There are two main complications. First, there is little capability for generating a large number of samples because the process to find the presence/absence of the microorganism in some cases may require hours or days and also it has a high cost (5–20€ per analysis) compared with acquiring data from a sensor (practically 0€). Second, there is high dimensionality because of the number of factors that may affect the presence of the microorganism. Moreover, the combination of these two drawbacks amounts to a statistical challenge when managing the samples to which the microbiological risk predictive model is to be applied. Therefore, we identified techniques that could function with a small number of samples and are studying the capacity for generating samples of a synthetic nature, in line with historical knowledge on the line and previous records. We are also analyzing techniques to reduce dimensionality in nonlinear problems in order to reduce the number of variables to be taken into account in the model. We are beginning with classic neural networks and plan to migrate to Back Propagation solutions which, according to the bibliography, have worked well, and we shall test new configurations using Reinforcement Learning techniques to support the lack of samples.

Results: At the time of writing, we do not yet have conclusive results allowing us to choose one technique rather than another.

5.3 Application of AI to the Extraction of Knowledge from Texts Referring to the Market and the Sector

In this third area of application of AI techniques in the food sector, we face a problem that falls outside production but is very relevant for food firms because it directly affects the detection of opportunities for innovation or of food safety risks based on monitoring the environment, in this case, the digital media environment.

The general goal is to extract knowledge published in digital media such as social media, professional publications, and generalist media that more or less specialize in the food sector. From the technological point of view, the goal is to classify and quantify texts and publications in certain categories that can add knowledge to certain incipient signs identified by an expert.

For this type of problem, any references on technical advances come from research. The sources on which these techniques are used in the sector are of various types: social media (Twitter), publications in specialized journals, blogs, publications by leading public agencies, etc. In all of them, natural language processing techniques are used. Since the objective is to establish deep learning models allowing for the classification and processing of captured texts, the first step is always correct preparation of the texts using techniques such as stemming (taking the semantic root of the words) and tokenization (dividing long texts into smaller units for processing). To create the models, word embedding techniques are used to allow terms and concepts to be represented as multi-dimensional vectors so that they can be properly processed by deep learning.

Some initiatives such as Injadat (2016) and Lee (2018) have identified working formulae for the classification of texts and publications, and neural network techniques (the basis of deep learning) and vector support machines have been identified as the most promising for texts coming from social media. Another interesting discipline for solving this type of problem is "opinion mining" (Pournarakis et al., 2017), that is, techniques to extract opinions from texts published by users and to associate them with the concepts that appear in such texts.

Another approach in the field of social media mining is to try to identify crises and threats from rumors. An initiative by Chen (2018) tries to identify rumors in social media not from a classification system approach but as an anomaly detection problem, by searching for changes in the pattern of communication of individual users.

Today the most usual solutions for this type of problem are grouped in 'ensembles', which combine various techniques, observing the results and providing a final solution for each problem. The most promising solutions that have obtained the best results in leading competitions in machine learning (https://www.kaggle.com/) combine various ensembles to achieve high-precision classification.

In the area of food, organizations such as the European Food Safety Agency (EFSA) and TNO have attempted to identify risks emerging in food from proof of concept using semantic technologies to identify relations between the appearance of key words in the abstracts of scientific articles and an ontology of food safety risks.

Today, two families of machine learning models are appearing for possible application to natural language problems, adopting a new approach and searching for the limits of modeling using revolutionary neural network techniques. One type of algorithm is Bidirectional Encoder Representations from Transformers (BERT, 2019), proposed by Google. This natural language model is a pre-trained model with a large amount of texts in English. It is unique in that it is based on unsupervised training with a Wikipedia text and 800 million words from a book base. It differs from traditional methods in which models such as Word2vec can generate a vector representation for each of the words it trains without taking into account the context (previous or subsequent words or sense of the text). For example, in English the word "right" would have the same representation in the sentence "*I'm sure I'm right*" as in "*Take a right turn.*" The main novelty is that BERT can represent the word "right" in its different senses because it creates the bidirectional concept, that is, the representation of the word takes into account the previous and subsequent words in the text, and this can be implemented by deep, bidirectional neural networks. The main advantage is that, based on this language training and with few samples, it can be adapted to a specific domain that we are interested in, such as discovering knowledge in the agrifood sector. The second main reference in this type of algorithm, which is also used in other contexts such as image recognition, is GPT, currently active in its versions GPT-2 and GPT-3. It is promoted by OpenAI (2019) whose principles take their inspiration from BERT, but it stands out for its capacity to predict the next word in a sentence or to complete words lost in a text, whatever the domain. The OpenAI team has released versions of this model but they are limited and go together with an open access article that describes the AI methodologies, mainly Transformers technology, which can establish relations between words following a given time line and an order when representing them.

5.3.1 Case Study of AI Applied to Supporting the Discovery of Emerging Risks in Food

Description of the problem: Specialists in the field of emerging risk detection in the food industry perform periodic reviews of hundreds of documents published in different sources to identify signs that might support an emerging risk hypothesis that they have identified a priori. This process sometimes leads the experts to analyze documents whose title seems to associate them with a specific risk but which turn out to be unrelated to that risk or any other, so reading them has been a waste of time. Therefore, to have a system that classifies incoming texts a priori into pre-set categories would save a lot of time and effort. Emerging risks are usually associated with regulatory trends on ingredients or products used by the industry, but may also be associated with risks that a product might be rejected in a market in relation to a reputational crisis of an ingredient or a product category, or even a negative association of a commercial brand in publications in social media.

Solution proposed by AINIA (see Fig. 3): Development of a computational model for text classification based on machine learning and on the Word Embeddings

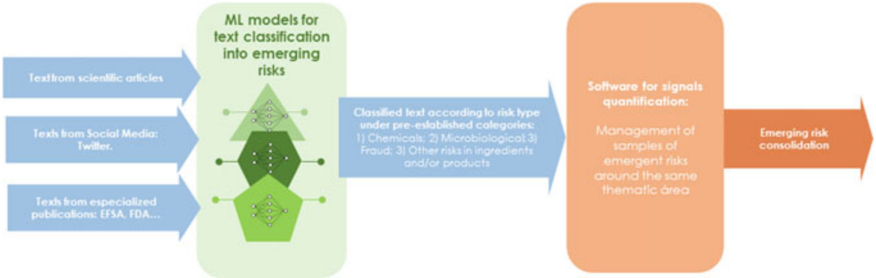


Fig. 3 Schema of ML models for text classification relating to emerging risks

methodology combined with neural networks. For this purpose, we trained more than 10,000 texts from about 500 different sources of articles and tagged them in four different categories of food risk (chemical, microbiological, fraud, ingredients, or products) plus one more for rejections, that is, to indicate that the text does not refer to a risk.

Results: The classifier achieved almost 95% precision in text classification. This reduced by half the time the experts devoted to identifying risks, minimizing the time to read articles not associated with a food risk. There is still room for improvement by increasing the number of samples in the model and adopting a system that will allow for periodic, semi-automatic updating. This aspect is to be covered in subsequent stages of the initiative.

6 Summary of the Main Advantages of the New Solutions, Benefits Obtained and Expected Future Benefits

6.1 Main Advantages

The areas of interest analyzed above focus on key activities in the food industry such as quality control in the production plant, food safety, promotion of product innovation, and analysis of threats for products. In all of them we have identified potential impacts and advantages for the sector, which can be summarized as follows:

– Improved perception of quality by consumers and customers of food firms: Adopting AI to establish links between production conditions and a product's target variables is key for standardizing optimal production conditions and promoting quality while adjusting aspects such as cost.
– Standardization of production conditions: AI can help detect the most favorable conditions so that they can be maintained or the least favorable so that they can be avoided.

- Waste reduction: AI can monitor production to minimize product rejection and to improve sustainability by saving the energy used on products that would not reach the market.
- Fewer quality complaints: When products are standard and within specifications, there are fewer quality complaints from customers and consumers. This avoids the cost of product withdrawals and a negative impact on the brand image.
- Fewer product safety crises: By monitoring the risk of microorganisms appearing that might impair quality or safety in both perishable and non-perishable food products in real time, AI allows for progress away from a reactive microbiological monitoring model following certain HACCP parameters to a model based on quantification, monitoring and management of risk.
- Improved decisions on product destination. AI can help in decisions about the destination of products that involve a potential risk. For example, risk can be classified as low, medium, or high, so products with an acceptable level of risk can be sent to markets where it would be legally permissible.
- Risk mitigation. AI techniques can identify the main causes behind a quality or food safety problem in a product or during its shelf-life. This allows for policies to be adopted that will mitigate the most relevant of these.
- Identification of innovation opportunities in line with market demands and with trends in the sector. AI can help filter, process, classify, and quantify information found in texts published in digital media. It can help identify product launches by other firms in similar product categories and can identify the degree of innovation of competitors and of the market in which a firm operates, in addition to many other analyses that can promote a firm's innovation strategy.
- Identification of threats to a firm's products that might have an impact on its strategic plans for innovation and product development. In opposition to the previous point, AI can identify the paths that a firm should not follow in its market strategy: avoiding unacceptable complaints from consumers or the market, avoiding specific ingredients or suppliers with a questionable reputation, and avoiding the effect of regulatory changes on products in the short and medium term, among many other aspects of risk monitoring.

In addition, there are other areas of the agrifood sector in which the application of AI might be of interest:

- Simulation of bioprocesses: These are techniques used to model advanced biological processes at human level (e.g. digestion models) and microbiological level (e.g. microbial fermentation processes as in wine and beer) or to model processes to create by-products such as biogas or proteins.
- Agronomic models for crop growth and production: Data-based machine learning models can be used to predict the volume of an agricultural product that a specific plot or farm can generate in the short, medium, or long term.
- Simulation of product demand: Algorithmic techniques are being devised that stress historic time series in order to predict the future behavior of a signal based on models that take into account this time component. For example, after eliminating

the effect of exceptional situations or anomalies, historic data on sales can be used to predict future demand for a product category.

7 Conclusions

In the application of AI to the food sector, the most important trends that will determine specific areas of research will focus on certain technical and technological challenges inherent in the nature of the problem, the context of use, the type of data and the capacity of producer firms to correctly sample data by following procedures and storing them in a structured fashion. The challenges identified for progress in the application of such technology to the industry are the following.

Complexity for aligning data samples: Firstly, in order to draw up a predictive model using deep learning techniques, it is essential to digitally characterize a product sample that has been through a set of production processes so that microbiological analysis can be performed or quality parameters monitored at different points in the production process and/or during the product's shelf-life. Total digital tracing of all the relevant parameters depends to a large degree on the preparation of the firm's information systems, and on the construction of software components that can align the data and link them for subsequent use. A lack of data or a rejected data sample may invalidate the data for a model. Also, this amounts to a major organizational and technological challenge for firms, especially for small or medium firms which are the majority in this sector. It is essential to work with production and quality experts to establish the logic for allocating certain process data to a specific sample.

Secondly, when modeling, the time component of, for example, a baking or curing process may affect not only the conditions of the current product but also which products were produced before or how long the process has been under way. The success of this phase of data sample alignment determines the success of the predictive model and requires effort and decision-making and, in some cases, will be based on trial and error. Aligning samples by process, product quality, or food safety variables may become a useful technological starting point for many other applications related to product improvement or other microbiological or quality risks for firms and may amount to a differentiating element and progress toward advanced use of the digitalized data.

Limited references on deep learning in dynamic industrial contexts. Many references to deep learning techniques are available in the research community but very few can be found in dynamic industrial contexts and practically none in the field of food safety. It is therefore a challenge to consolidate certain cutting-edge techniques that might be appropriate for solving common problems and achieve acceptable levels of success in prediction (prognosis) and in simulation of production processes. There is not yet a consensus in this regard among the research community either for industrial sectors in general or for the food sector in particular. This means that the adoption of AI in industry is still under-developed in comparison with other methods such as image processing or natural language.

Money and time constraints for collecting and processing good-quality microbiological samples. Microbiological information and much objective data on quality parameters come from laboratory analysis so require sample processing, reagents, and time to analyze each microorganism or quality parameter. While obtaining process data is relatively cheap because of the use of sensors and actuators on production lines, it is not so easy or so cheap to automatically obtain information on quality variables. This drawback is inherent in the development of specific models and has to be resolved using strategies based on generalist models and on specialization using small data samples in a specific product category.

"Imbalanced classes" (Krawczyk, 2016): In a food production environment, most of the products are usually sold with good microbiological and product quality within the specifications. Obviously, products with microbiological spoilage or with impaired quality may appear occasionally, but this is rare. In the case of pathogens, this scenario is especially acute because production processes are designed to minimize them, although such elimination might be less effective in fresh products. Therefore, if we set up a data sample collection process in a production environment, most of the samples captured will meet specifications so will be of little interest for helping to solve problems. Also, there will be few samples and the data will not be very representative of problems that might arise on the line. This problem is known as "imbalanced classes" and there are techniques that can correct it under certain conditions. However, more work is needed on them as this is a real problem that will become more important in the future.

Multivariance with large numbers of variables. The many variables considered may include categorical, non-numerical variables such as ingredients, suppliers and other types of product information that may have an influence on the shelf-life of products. When developing predictive models, there should be a balance between the number of variables that describe the data in a sample, and the number of samples available for modeling. Predictive models will be simpler and more likely to show good predictive behavior if they include a small number of variables that are very significant in the process. In order to achieve success in this context, multidisciplinary teams are needed to identify the variables and parameters that are really causing the phenomenon to be modeled, avoiding signals that might interfere with the technological capability of the models for describing the reality.

Explainability of predictive models: It is currently being found in the AI scientific community that techniques such as neural, or black box, techniques do not afford good explainability of predictive results nor of what happens internally in the decision-making processes of the neural network. That is, if decisions are taken based on AI algorithms, it is not fully understood exactly why certain results are reached. This is what is known as Explainable Artificial Intelligence (XAI) (Arrieta et al., 2020). It is very important to work on the explainability of models in order to help resolve the problem of mitigating microbiological risk in order to learn more about the most effective opportunities for risk mitigation. It is also important from an ethical point of view because there is debate on whether what algorithms can process and present can be believed because, if there is any bias in the data, any predictive algorithm will include the same bias. Such a bias would exist, for example, if, when analyzing social media, we only take into account opinions on certain brands and products that are not available to all, omitting a certain group. This would result in an erroneous strategy if, for example, that group is the usual customer base for the product being studied. Work is also being done on grey box models which reinforce black box knowledge with data based on real expert knowledge. This results in a more powerful working base, overcoming to some extent any polarity introduced by certain data groups.

References

Abbasimehr, H., Shabani, M., & Yousefi, M. (2020). An optimized model using LSTM network for demand forecasting. *Computers & Industrial Engineering, 143*, 106435.

Arrieta, A. B., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

BERT. (2019). Understanding BERT: Is it a Game Changer in NLP? Available at: https://towardsdatascience.com/understandingbert-is-it-a-game-changer-in-nlp-7cca943cf3ad.

Chen, W., Zhang, Y., Yeo, C. K., Lau, C. T., & Lee, B. S. (2018). Unsupervised rumor detection based on users' behaviors using neural networks. Pattern *Recognition Letters, 105*, 226-233.

Data & Trends EU Food And Drinks Industry, from Food Drink Europe. http://www.fooddrinkeurope.eu/uploads/publications_documents/DataandTrends_Report_2017.pdf.

Deng, Y., et al. (2019). Prediction model of PSO-BP neural network on coliform amount in special food. *Saudi Journal of Biological Sciences, 26*, 1154–1160.

Dilipkumar, D. (2020). States, Actions, Rewards—The Intuition behind Reinforcement Learning. Available at: https://towardsdatascience.com/states-actions-rewards-the-intuition-behind-reinforcement-learning-33d4aa2bbfaa.

European Parliament. (2015). *Industry. Digitalisation for productivity and growth.*

European Union. (2021). Food Safety in the EU. Retrieved June, 2021, from https://europa.eu/european-union/topics/foodsafety_en.

FoodDrinkEurope. (2020). Data & Trends EU Food And Drinks Industry. Available at: https://www.fooddrinkeurope.eu/wpcontent/uploads/2021/02/FoodDrinkEurope-Data-Trends-2020-digital.pdf.

Food & Drug Administration. (2020). *How to start a food business*. Retrieved June 2021, from https://www.fda.gov/food/foodindustry/how-start-food-business.

Fernández-Navarro, F., et al. (2010). Development of a multi-classification neural network model to determine the microbial growth/no growth interface. *International Journal of Food Microbiology, 141*, 203–212.

Gosukonda, R., et al. (2015). Application of artificial neural network to predict Escherichia coli O157:H7 inactivation on beef surfaces. *Food Control, 47*, 606–614.

Geng, Z., Zhao, S., Tao, G., & Han, Y. (2017). Early warning modeling and analysis based on analytic hierarchy process integrated extreme learning machine (AHP-ELM): Application to food safety. *Food Control, 78*, 33–42.

Geng, Z., Shang, D., Hana, Y., & Zhong, Y. (2019). Early warning modeling and analysis based on a deep radial basis function neural network integrating an analytic hierarchy process: A case study for food safety. *Food Control, 96*, 329–342.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Holley, R. A., & Patel, D. (2005). Improvement in shelf-life and safety of perishable foods by plant essential oils and smoke antimicrobials. *Food Microbiology, 22*(4), 273–292.

Injadat, M. N., & FadiSal, A. N. (2016). Data mining techniques in social media: A survey. *Neurocomputing, 214*, 654–670.

Jiang, B., He, J., Yang, S., Fu, H., Li, T., Song, H., & He, D. (2019). Fusion of machine vision technology and alexnet-CNNs deep learning network for the detection of postharvest apple pesticide residues. *Artificial Intelligence in Agriculture, 1*, 1–8.

Kaggle. (2017). VGG-19. VGG-19 Pre-trained Model for Keras. Available at: https://www.kaggle.com/keras/vgg19/home.

Kaiyi, W., Chunjiang, Z., Tian, Z. F., Juan, W. X., Hong, C., & Fang, Y. Y. (2007). Application of a support vector machine to HACCP in the animal meat industry. *New Zealand Journal of Agricultural Research, 50*(5), 743–748.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence,* 221–232.

Lee, I. (2018). Social media analytics for enterprises: Typology, methods, and processes. *Business Horizons, 61*, 199–210.

Melo, F. S. (2001) *Q-learning: Convergence of Q-learning: A simple proof*. Institute for Systems and Robotics, Instituto Superior Técnico, Lisboa, Portugal. Available at: http://users.isr.ist.utl.pt/~mtjspaan/readingGroup/ProofQlearning.pdf.

MIT Technology Review. (2020). Why asking an AI to explain itself can make things worse. Available at: https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explain-itself-can-make-things-worse/

Mortimore, S. (2000). An example of some procedures used to assess HACCP systems within the food manufacturing industry. *Food Control, 11*(5), 403–413.

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response surface methodology: Process and product optimization using designed experiments.* John Wiley & Sons.

Mikolov, T., Karafiát, M., Burget, L., & Černocký, J. (2010). Recurrent neural network-based language model. In *Eleventh Annual Conference of the International Speech Communication Association.*

Nguyen, T. T., Nguyen, N. D., Vamplew, P., Nahavandi, S., Dazeley, R., & Lim, C. P. (2020). A multi-objective deep reinforcement learning framework. *Engineering Applications of Artificial Intelligence, 96*, 103915.

Oladunjoye, A. O., Oyewole, S. A., Singh, S., & Ijabadeniyi, O. A. (2017). Prediction of Listeria monocytogenes ATCC 7644 growth on fresh-cut produce treated with bacteriophage and sucrose monolaurate by using artificial neural network. *LWT—Food Science and Technology, 76*, 9–17.

OpenAI. (2019). Better Language Models and Their Implications. Available at: https://openai.com/blog/better-language-models/.

Pan, L., Pouyanfar, S., Chen, H., Qin, J., & Chen, S. C. (2017). Deepfood: Automatic multiclass classification of food ingredients using deep learning. In: *2017 IEEE 3rd International Conference on Collaboration and Internet Computing, CIC* (pp. 181–189). IEEE.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345.

Perez, J., Marinkovi, J., & Barcel, P. (2019). *On the turing completeness of modern neural network architectures.* Department of Computer Science, Universidad de Chile & IMFD Chile.

Pournarakis, D. E., Sotiropoulos, D. N., & Giaglis, G. M. (2017). A computational model for mining consumer perceptions in social media. *Decision Support Systems, 93*, 98–110.

Reinforcement Learning—An Introduction, a book by the father of Reinforcement Learning-Richard Sutton and his doctoral advisor Andrew Barto. An online draft of the book is available here http://incompleteideas.net/book/the-book-2nd.html.

Sartal, A., Martínez-Senra, A. I., & García, J. M. (2017). Balancing offshoring and agility in the apparel industry: Lessons from benetton and inditex. *Fibres & Textiles in Eastern Europe.*

Sartal, A., Carou, D., Dorado-Vicente, R., & Mandayo, L. (2019). Facing the challenges of the food industry: Might additive manufacturing be the answer?. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 233*(8), 1902-1906.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning—An Introduction. MIT Press Cambridge

Tensorflow. (2022). Word2vec family of algorithms. Available at: https://www.tensorflow.org/tutorials/text/word2vec.

Verboven, P., et al. (2020). Digital twins of food process operations: The next step for food process models? *Current Opinion in Food Science, 35*, 79–87.

Word2vec family of algorithms. https://www.tensorflow.org/tutorials/text/word2vec.

Xia, M., et al. (2020). Data-driven prognosis method using hybrid deep recurrent neural network. *Applied Soft Computing Journal, 93*, 106351.

Yan, R., et al. (2019). Deep learning for fault diagnosis and prognosis in manufacturing systems. *Computers in Industry, 110*, 1–2.

Machine Learning for the Prediction of Edge Cracking in Sheet Metal Forming Processes

Armando E. Marques, Pedro A. Prates, Ana R. Fonseca, Marta C. Oliveira, Martinho S. Soares, José V. Fernandes, and Bernardete M. Ribeiro

Abstract This work aims to evaluate the performance of various machine learning algorithms in the prediction of metal forming defects, particularly the occurrence of edge cracking. To this end, seven different single classifiers and two types of ensemble models (majority voting and stacking) were used to make predictions, based on a dataset generated from the results of two types of mechanical tests: the uniaxial tensile test and the hole expansion test. The performance evaluation was based on four metrics: accuracy, recall, precision and F-score, with the F-score being considered the most relevant. The best performances were achieved by the majority voting models. The ROC curve of a majority voting model was also evaluated, in order to confirm the predictive capabilities of the model. Globally, ML algorithms are able to predict the occurrence of edge cracking satisfactorily.

1 Identification of the Sector

Sheet metal forming is a manufacturing technique capable of producing a high volume of components at a relatively low cost. Thus, this technique is widely used in the automotive and aerospace industries. The automotive industry is characterized by

A. E. Marques · P. A. Prates · M. C. Oliveira · J. V. Fernandes
Department of Mechanical Engineering, CEMMPRE, University of Coimbra, 3030-788 Coimbra, Portugal

P. A. Prates (✉)
Department of Mechanical Engineering, Centre of Mechanical Technology and Automation (TEMA), University of Aveiro, 3810-193 Aveiro, Portugal
e-mail: prates@ua.pt

A. R. Fonseca · M. S. Soares
Toolpresse, Peças Metálicas por Prensagem Lda., Parque Industrial De Vendas Novas, 7080-341 Vendas Novas, Portugal

B. M. Ribeiro
Department of Informatics Engineering, University of Coimbra, CISUC, 3030-290 Coimbra, Portugal

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8_6

a constant need for innovation, in order to be able to guarantee profits while ensuring that their products respect quality, safety and environmental impact demands. This industry has great relevance in the European market, and in particularly in Portugal where there are over 400 companies dedicated to the production of components, of which about 8 percent of business volume derives from metallic components.

2 Problem Statement

High strength steels are frequently used in components produced by sheet metal forming processes, particularly in the automotive industry, as their use allows for the reduction of vehicle's mass, while still guarantying good mechanical properties. However, when producing components with these materials, the occurrence of edge cracking is relatively common, particularly when they have bending regions where the ratio between curvature radius and the sheet thickness is low. This edge cracking consists of the occurrence of fractures in a component, usually at the outer edge of a bent area, where the strain path corresponds to uniaxial tension.

3 Previous Solutions

The traditional approach to process design was based on trial-and-error. This approach is costly and time-consuming, making it unfeasible in today's extremely competitive market. As such, researchers are looking for more efficient alternative methods to apply to process design. At first, the focus was on the application of the finite element method (FEM) for the simulation of forming processes. However, FEM simulations can be computationally expensive when applied to complex forming processes, and a significant number of simulations can be required to obtain good design solutions. Another limitation of the FEM solution is the fact that material variability is not properly taken into consideration. In fact, there can be significant differences in mechanical behaviour between the various metal sheet coils received during production. This difference in mechanical behaviour leads to the occurrence of defects when using some of these material coils during production, which was not predictable during the design process.

4 Description of the Proposed Solution

In order to reduce the scrap rate associated with the material mechanical behaviour variability mentioned previously, some form of material evaluation should be performed on newly received materials, before production, to confirm whether or not each material has the mechanical properties necessary to avoid the occurrence of defects.

To the author's knowledge, the application of ML algorithms in this type of evaluation is not common, however, these algorithms have proved effective in other applications related to both forming process analysis and material parameter identification. Among the various ML algorithms available, artificial neural networks (ANN) are the most used. Aghasafari et al. (2014) applied ANN to a hot rolling process in order to predict flow stress variations. Many combinations of training algorithms and transfer functions were evaluated in this work, in which the predictive performances obtained by the ANN built with the best combination (Levenberg-Marquardt training algorithm and tan-sigmoidal transfer function) surpassed those obtained by a conventional least-squares method. Hartmann et al. (2019) used an ANN to generate tool paths for an incremental sheet metal forming process, using data representing the desired piece geometry as inputs, and the results showed potential for practical applications. Spathopoulos et al. (2020) tested the performance of an ANN in the springback prediction of a S-Rail forming process, achieving a promising mean-square error value. Merayo et al. (Fernández et al., 2020) and Koenuma et al. (2020) applied ANN for predicting the stress-strain curves of aluminium alloys. The first work focuses on estimating four material parameters, namely, Young's modulus, yield strength, ultimate tensile strength and elongation at break, and then on defining a bilinear approximation of the real stress-strain curve. This approach was based on a relatively large dataset (2000 alloys) and achieves good overall predictions for the material properties, while showing the difficulties that metamodels can have in making predictions for outliers. The second work used crystallographic texture data as input, and considered the stress-strain curve and r-values as outputs. Beskopylny et al. (Beskopylny et al., 2020) developed an ANN classification algorithm to group various steel grades, in regard to strength features, achieving a 95% predictive accuracy. Masi et al. (2021) developed a new ANN approach, called thermodynamics-based neural networks (TANN), consisting on an ANN in which the two basic laws of thermodynamics were encoded directly in the algorithm. This approach guarantees that any prediction remains thermodynamically consistent and proved to be more effective in the prediction of stress, energy and dissipation than standard ANNs. Gorji and Mohr (2019) implemented ANN to describe a constitutive material model and applied it to the simulation of tensile tests, obtaining results comparable to those generated by a J_2 plasticity model, which highlights the potential of ANNs to replace conventional models in finite element applications. While it is clear that ANN-based algorithms can achieve very positive results, the capabilities of other algorithms should not be discarded.

In the case of classification problems, alternative classification algorithms have been used in various areas, achieving results that can clearly compete with those of ANNs. Guevara et al. (2020) applied ML classification to fog computing applications, including various ANN configurations, decision trees (DT) and support vector machines (SVM). These algorithms were tested for various degrees of noise in the dataset, and DT proved to be the most robust for this application. Cervantes et al.

(2020) present a study on SVM, highlighting the various applications of these algorithms, including areas like image identification, face detection, written character recognition and some biology applications. Konstantopoulos et al. (2020) applied classification algorithms, including ANNs, SVM and random forest (RF), to the identification of relations between the structure and mechanical properties of carbon fibre reinforced polymers. In this case, SVM achieved superior results, particularly for the final validation dataset, although it is worth mentioning that this dataset was small, which led the authors to consider that overfitting may have occurred in the remaining models. Yucalar et al. (2020) applied ML classification to the prediction of software defects during the development process. This work focused on the application of ensemble models, including 10 different types of ensemble predictors and 16 different single classifiers as base predictors. This wide variety of analysed classification algorithms highlights the amount of algorithm options potentially available and the need to evaluate these various options to make an informed decision, in terms of which one to use for a certain application.

The current work consists of evaluating the performance of various ML single classification algorithms, as well as of ensemble models, in predicting the occurrence of edge cracking in sheet metal forming processes. The training and performance evaluation processes are based on a dataset generated from experimental results of two types of mechanical tests, the uniaxial tensile test and the hole expansion test, with both showing a strain path of uniaxial tension, similarly to the regions of the component where edge cracking generally occurs.

4.1 Dataset Generation

The success of the learning stage of an ML algorithm depends on the creation of a good dataset. A dataset consists of the collection of all available information about the problem in question, upon which the algorithm training process is based. The dataset used in this work contains experimental results obtained from two different mechanical tests: the uniaxial tensile test and the hole expansion test. These tests were performed for samples obtained from a total of 176 different sheet metal coils, which for the purposes of this work are all considered to be of different materials, which translate to 176 entries in the dataset.

Two uniaxial tensile tests were performed for each coil, one for the rolling direction and the other for an angle of 90º with the rolling direction. The results obtained from these tests are the yield strength (R_e), the ultimate tensile strength (R_m) and the corresponding strain value (ε_{Rm}), and the percent elongation at fracture for initial gauge lengths of 50 mm (E_{50}) and 95 mm (E_{95}). These results correspond to the input variables of the dataset, and the minimum, mean and maximum values obtained for each variable are shown in Table 1.

The samples used for the hole expansion tests had an initial hole diameter of 20 mm, and the test was performed up to punch displacement values that guaranteed the fracture of all samples. The result obtained from these tests is the strain at fracture.

Table 1 Distribution of the input variables

	$R_{e,0°}$ [MPa]	$R_{e,90°}$ [MPa]	$R_{m,0°}$ [MPa]	$R_{m,90°}$ [MPa]	$\varepsilon_{Rm,0°}$	$\varepsilon_{Rm,90°}$	$E_{50,0°}$ (%)	$E_{50,90°}$ (%)	$E_{95,0°}$ (%)	$E_{95,90°}$ (%)
Minimum	138.56	137.30	257.10	261.99	0.070	0.008	9.0	11.0	9.54	10.53
Mean	349.31	359.45	424.68	425.74	0.153	0.143	28.70	28.04	23.96	23.35
Maximum	577.93	630.09	615.63	675.04	0.251	0.260	51.0	52.0	40.89	40.65

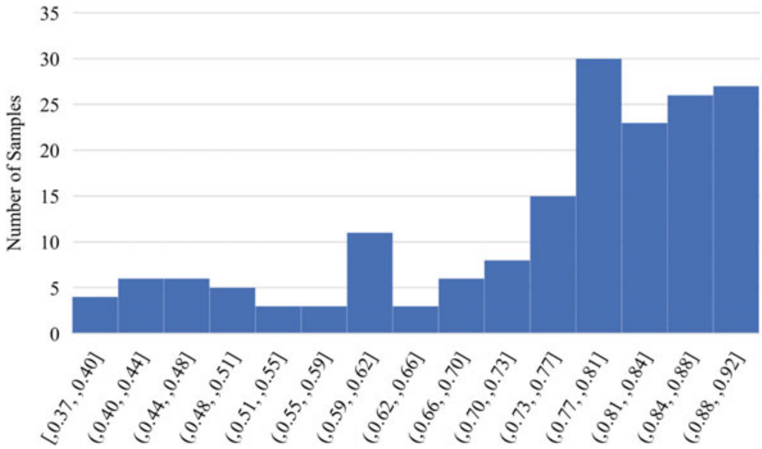


Fig. 1 Distribution of the strain values at fracture obtained for the hole expansion tests

Its values allow to establish a criterion to define whether edge cracking occurs or not. This criterion is established based on a target strain value, which depends on the component under study. If the strain at the moment of fracture for a specific material is lower than the target strain, then edge cracking will occur for this material. The adoption of this criterion is justified by the fact that a uniaxial tension strain path occurs both at the edge of the hole, during a hole expansion test, and in the critical zones of the components where edge cracking typically occurs. As such, a correlation can be established between the strain values in the critical zones of the component and those achieved in the hole expansion test. Two different target values were considered in this work. The first is a strain value of 0.725, which corresponds to the average of the strain values obtained for the 176 different materials (these values vary between 0.34 and 0.91, and their distribution is represented in Fig. 1). For this target value, 62 of the 176 materials lead to edge cracking. Since this is a relatively low amount, a second target value of 0.82 was considered, with 123 of the 176 materials leading to edge cracking.

4.2 Single Classifiers

The single classifier algorithms considered in this work are

- Multilayer Perceptron (MLP)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- Random Forest (RF)
- Logistic Regression (LR)
- Naïve Bayes (NB)

- k-Nearest Neighbours (kNN).

The models were built using python v2.7.18, with the SciKit-learn library (Pedregosa et al., 2011). The theoretical basis of the algorithms is briefly explained in the following sections (Dib et al., 2020).

4.2.1 Multilayer Perceptron

The multilayer perceptron (MLP) is a feed-forward neural network, consisting of one input layer with a number of nodes (neurons) equal to the number of input variables, an arbitrary number of hidden layers (minimum one) and one output layer. Each node is connected to the nodes of the next layer, but there is no connection between the nodes in the same layer. Each of the nodes of the hidden layers has a nonlinear function called the activation function. The output of a node in a hidden layer or in the output layer is given by the following equation:

$$z_i = \phi\left(\sum_j w_{ij} z'_j + b_i\right), \tag{1}$$

where z_i is the output of the current node i, z'_j is the value obtained from node j of the previous layer, w_{ij} is the weight associated to node i and z'_j, ϕ is the activation function and b_i is the bias term.

4.2.2 Support Vector Machines (SVM)

The support vector machines (SVM) are discriminative classifiers, which find the optimal separating hyperplane for the training data points, $\boldsymbol{x}$. It consists of identifying the hyperplane that maximizes the distance between itself and any datapoint corresponding to each class. This hyperplane is defined by

$$(\boldsymbol{w}^T \boldsymbol{x} - b) = 0, \tag{2}$$

where $\mathbf{w}$ is the normal vector to the hyperplane. Maximizing the distance between the hyperplane and the datapoints of each class means minimizing $\|\mathbf{w}\|$. In order to apply SVM to problems that are not linearly separable, a kernel trick is applied. This consists of transforming the problem space into an implicit, high-dimensional feature space, where linear separation is possible.

4.2.3 Decision Tree (DT)

A decision tree (DT) consists of a nonparametric classifier that uses simple rules to continuously split data. These rules are chosen so that information gain is maximized, which means minimizing the following function, F:

$$F = \sum_{i=1}^{m} \frac{n_i}{N_i} H(D_i), \tag{3}$$

where N is the number of examples in the resulting node i and n is the number of examples with the desired label in the mentioned node. H corresponds to an impurity function, such as entropy:

$$H(D) = -\sum_{i=1}^{m} p_i \log(p_i), \tag{4}$$

where p is the probability that a given example in the dataset corresponds to label i. The splitting process is repeated until all the samples present in each of the final nodes correspond to the same label, or until an early stopping criterion is satisfied.

4.2.4 Random Forest (RF)

The random forest (RF) consists of a combination of several decision trees, which are randomly generated and trained with different parts of the training set. In a classification problem, the predictions from the random forest correspond to the predictions made by more than half of the decision trees.

4.2.5 Logistic Regression (LR)

The logistic regression (LR) consists in fitting a logistic curve, which is a sigmoid curve, to the relationship between the input variables of the training set, x, and the corresponding labels, y. The logistic function is defined as follows:

$$y = \frac{1}{1 + e^{-x}}. \tag{5}$$

This formula can be extended with coefficients α and β, where α is the intercept of y and β is a regression coefficient:

$$y = \frac{1}{1 + e^{-(\alpha + \beta x)}}. \tag{6}$$

4.2.6 Naïve Bayes (NB)

Naïve Bayes is a classifier based on the Bayes theorem, under the (Naïve) assumption that every pair of input variables is independent. With the Naïve assumption, Bayes' theorem is represented by the following expression:

$$P(y|x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, \ldots, x_n)}. \quad (7)$$

Since, for a given dataset, the denominator will be the same for all entries, a proportionality can be considered:

$$P(y|x_1, \ldots, x_n) \propto P(y)\prod_{i=1}^{n} P(y). \quad (8)$$

The predictions are made based on the label that presents the highest probability:

$$y = \text{argmax}_y P(y)\prod_{i=1}^{n} P(x_i|y). \quad (9)$$

4.2.7 k-nearest Neighbours (kNN)

The *k*-nearest neighbours classifier does not create a model with the training data. Instead, the training data is simply recorded and, each time the model makes a prediction, the distance between each of the training points and the point for which the prediction will be made is calculated. Then, the *k*-nearest training points are selected and only these are considered for the prediction. The prediction can simply be the label corresponding to more than half of the training points considered, or weights can be given to each training point, so that from the *k* training points, the nearest ones are more influential.

4.3 Ensemble Models

The ensemble models combine various single classifiers, called base learners in this context, to make predictions. In theory, when combining various classifiers, these models are able of making better predictions, since they have reduced bias when compared with the single classifiers.

Two different types of ensemble models are considered in this work: majority voting and stacking models. In majority voting models, each base learner is trained individually and then is used to make predictions. The class with more than half of the votes is the one predicted by the ensemble model. Stacking models have two levels of learners. First, the base learners are trained and make predictions. Then, a different learner, called a meta-learner, is trained with the predictions of the base learners, and the meta-learner makes the final prediction of the ensemble model.

4.4 Model Calibration, Training and Evaluation

The calibration, training and evaluation processes of the models require a split of the dataset into a training set, which, as the name suggests, is used to train the models, and a testing set, for which the models make predictions after training. The training of the model consists on adjusting the model parameters, through an optimization algorithm, so that the predictions made by the model for the training set entries match as closely as possible to their real classes.

To enhance the performance of each model, the hyperparameters of each algorithm were calibrated through a trial-and-error process. This is an iterative process, in which a part of the training set is taken out of it to form a validation set. The models are first trained with the reduced training set, make predictions for the validation set and their performance is evaluated; then, hyperparameters are altered to try to improve the results. Once a satisfactory set of hyperparameters is found, the full training set is used in the training of the model and its performance is evaluated for the testing set.

The performance evaluation of the classification models is carried out by means of performance metrics, derived from the confusion matrix. A confusion matrix reports the predictions made by a model with respect to the real classes. For a problem with two classes, the confusion matrix has four elements, as shown in Table 2. The performance metrics considered for this work are accuracy, recall, precision and F-score. Accuracy represents the rate of correct model predictions from the overall predictions. Recall is the rate of positive cases in the testing set that the model correctly identified as positive. Precision is the rate of positive predictions made by the model that is correct. F-score is a more general performance metric that encompasses both recall and precision. It consists of the harmonic mean of these two metrics. The performance metrics are calculated as shown in equations 10-13.

Table 2 Confusion matrix of a problem with two classes

		Real class	
		1	0
Predicted class	1	True Positive (TP)	False Positive (FP)
	0	False Negative (FN)	True Negative (TN)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{10}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

For the problem addressed in this work, the correct predictions of positive cases (occurrence of edge cracking) are considered a priority and, therefore, more important. Typically, edge cracking occurs for a relatively small portion of the different coils, used during production. This means the dataset is highly unbalanced. This means that a model with low capacity to detect positive cases can still achieve high accuracy, by correctly predicting the negative cases. Thus, accuracy should be considered a support metric, and the performance metrics related to the detection of edge cracking, particularly the F-score, are considered more relevant when evaluating each model.

In order to obtain a more robust performance evaluation, 30 different combinations of training and testing sets were generated randomly from the dataset; for each combination, the training sets contain 70% of the dataset and the testing sets contain the remaining 30%. Performance metrics were calculated for each combination, in order to obtain the mean and standard deviation of each metric, for each ML algorithm.

After the performance of the various models was evaluated, a receiver operating characteristic (ROC) curve was generated for one of the models. This curve consists of a graph that plots the true positive rate (recall) as a function of the false positive rate, calculated as

$$\text{False positive rate} = \frac{\text{FP}}{\text{TP} + \text{FP}}. \tag{14}$$

The ROC curve is used to evaluate the sensitivity of the model to different thresholds between positive and negative cases, which determine the number of positive and negative cases in the dataset. The generation of this curve required the application of a criterion considering multiple values of target strain, in addition to those considered for the remaining analysis.

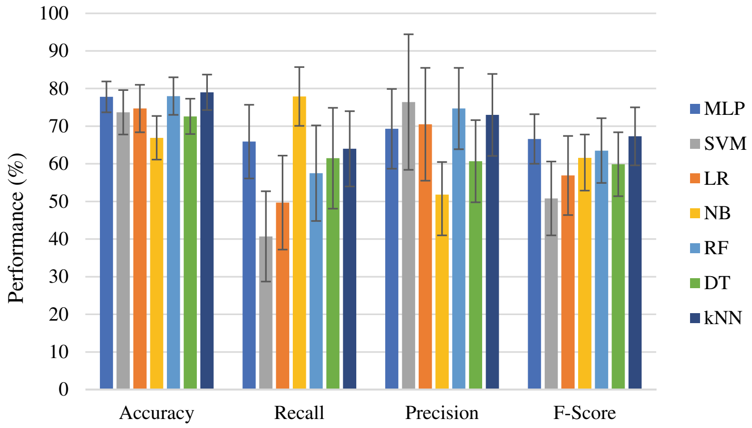


Fig. 2 Performance metrics for the single classifiers considering a target strain value of 0.725

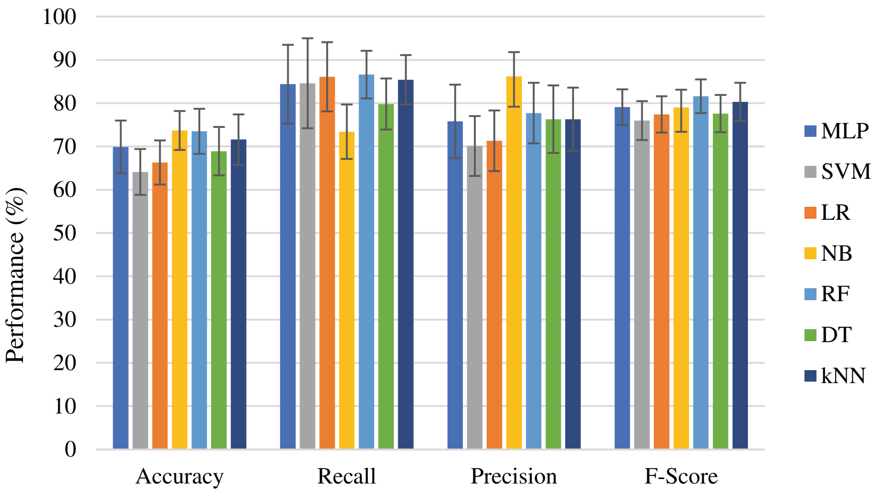


Fig. 3 Performance metrics for the single classifiers considering a target strain value of 0.82

4.5 Results and Discussion

Figures 2 and 3 present the performance metrics obtained for the single classifier models. These correspond to the cases with target strain values of 0.725 and 0.82, respectively.

For the target strain value of 0.725, accuracy is in general the highest metric, with mean values between 70 and 80% for all classifiers, except NB. The remaining metrics, which focus on the prediction of positive cases, show lower mean values and higher standard deviations. These results can be explained by the low number of positive cases present in the dataset, which means less informative data available on positive cases. The low number of real positive cases also makes each wrong prediction of a positive case have more impact on recall, as it represents a higher percentage of the total positive cases present in the training set. The results obtained for the target strain value of 0.82, which leads to significantly more real positive cases, confirm these conclusions, showing in general better average values and lower standard deviations for recall, precision and F-score, but slightly lower accuracy.

There is significantly greater variation between the performance metrics of each classifier for the target strain value of 0.725. Considering the fact that, as previously stated, positive cases (edge cracking) tend to be the exception and not the rule for the problem in question, the target strain value of 0.725 is considered more relevant in determining which classifiers offer the best performances. Thus, based on the F-score values, MLP and kNN single classifiers are considered to have the best performance for this problem with, respectively, 66.6% and 67.3%.

The performance metrics obtained for the majority voting ensemble models are shown in Figs. 4 and 5. The labels on the right-hand side of the figures depict which single classifiers are included as base learners in each of the models.

When compared with the single classifiers, the majority voting models perform better overall. For the target strain value of 0.725, they have better average values for accuracy (all approximately 80%, which among the single classifiers, only kNN, with 79 %, can compete with) and precision (the highest precision average value achieved by a majority voting model is 80.1%, while the best value obtained by a single classifier, SVM in this case, is 74.7%). The majority voting models with the highest average F-score values also compete with the best-performing single classifiers. The highest average F-Score achieved in this case is 68.7%, which represents an increase of 1.4% when compared to the kNN single classifier. When considering the strain target value of 0.82, the majority voting models achieve mean F-scores of 80%, a value achieved only by the RF and kNN single classifiers, while having slightly superior accuracy, with the highest average value achieved by a majority voting

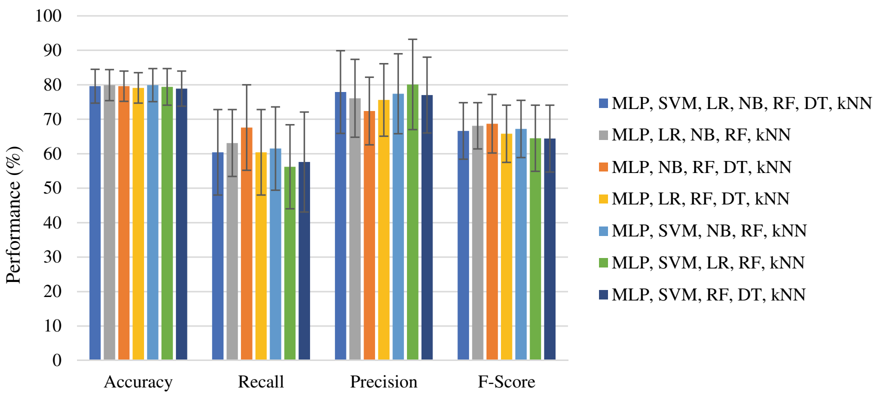


Fig. 4 Performance metrics for the majority voting models, with a target strain value of 0.725

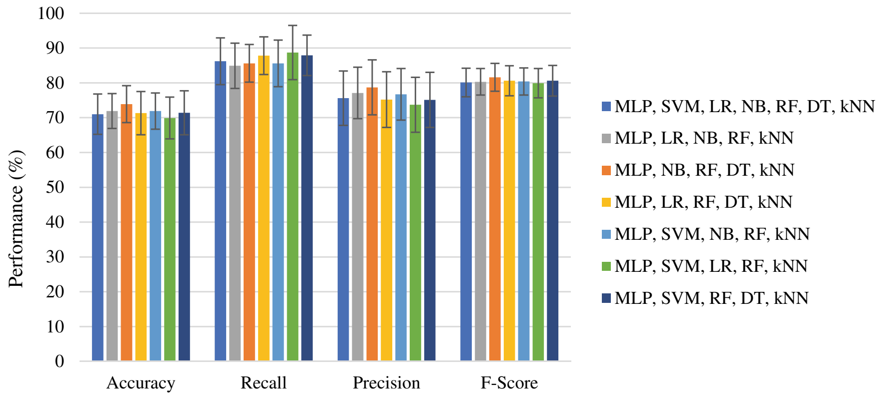


Fig. 5 Performance metrics for the majority voting models, with a target strain value of 0.82

model being 73.9%, while the RF and kNN single classifiers achieved 73.5% and 71.6 %, respectively. It is also worth noting that there is much lower variation in performance metrics between the various majority voting models, when compared to the single classifiers.

Figures 6 and 7 show the performance metrics obtained for the stacking models. The labels on the right side of the figures identify the single classifier used as meta-learner for each model. The base learners used in each model are the remaining six single classifiers considered in this work. The performance of the stacking models is competitive with the remaining models when considering a target strain value of 0.82, although with slightly higher standard deviations. However, for the target

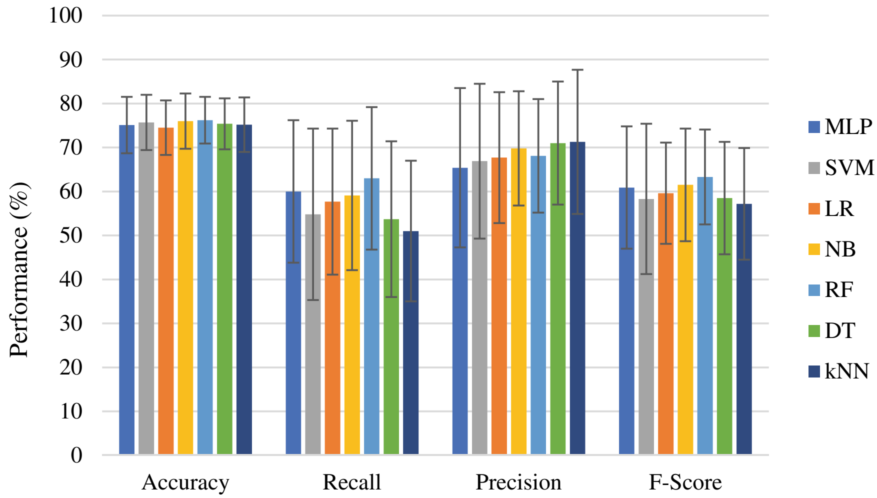


Fig. 6 Performance metrics for the stacking models considering a target strain value of 0.725

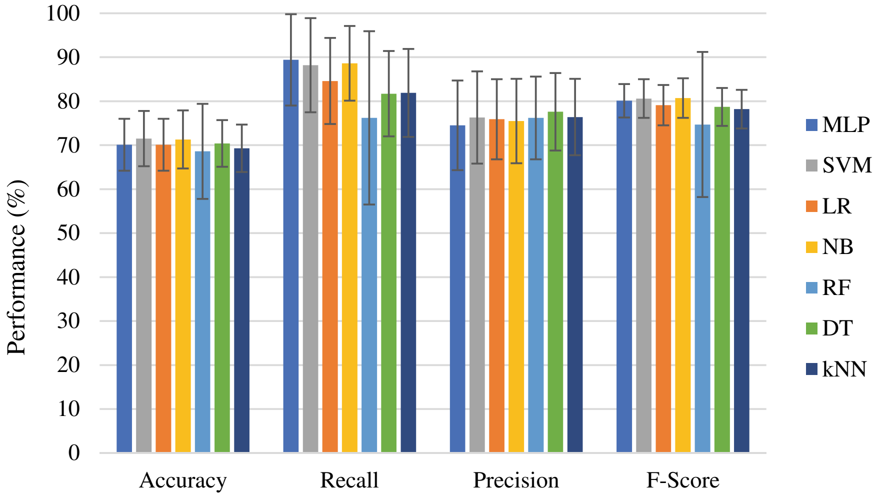


Fig. 7 Performance metrics for the stacking models considering a target strain value of 0.82

strain value of 0.725, these models perform clearly worse than the majority voting models and the best single classifiers, with lower mean values for all metrics and significantly higher standard deviations. For instance, the average F-score values achieved by these models vary between 58.3 and 63.3%, with standard deviations between 10.8 and 17.1%, while the majority voting models achieve average F-score values between 64.4 and 68.7%, with standard deviations between 6.7 and 9.7%. Thus, the stacking models are not considered effective or reliable for application to the current problem.

Overall, the majority voting models are considered the best choice for application to the problem under analysis, although exceptions can be made for the single classifiers with best performance, since they also present good performances, while requiring significantly less computational power, which can be a relevant factor for cases with much larger datasets. The performance differences between the various majority voting models are relatively small, so the application of any one of them can be recommended.

In order to better understand how the models perform for various different target strain values, a ROC curve was generated for the majority voting model that included all seven single classifiers as base learners. This curve is represented in Fig. 8. Each point of the curve corresponds to a different target strain value, with values ranging between 0.58 and 0.83 being considered. The red line at 45° represents a model that is as effective as a random guess, similar to a coin toss. The ROC curve of a useful classification model will be above this line, and the distance between the two is directly proportional to the quality of the predictions. Another way of relating the ROC curve to the 45° line is through the area under the curve (AUC). The AUC can be used to assess the robustness of a model. A higher AUC means a greater capacity to distinguish positive from negative cases. The 45° line has an AUC of 50%, while the

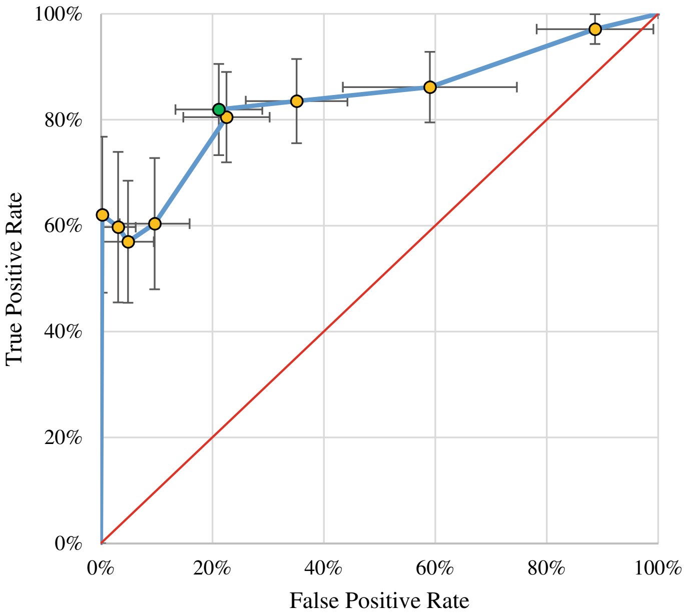


Fig. 8 ROC curve for the majority voting model that includes all seven single classifiers

current ROC curve has an AUC of approximately 85%, which confirms the quality of the model under analysis. The point identified in green corresponds to the target strain value of 0.762, which is close to the median of the strain values at the moment of fracture in the hole expansion test, and represents the best performance of the model among the various target values considered, with a true positive rate of 82% and a false positive rate of 21%. In practice, this means that the model will achieve better performance the closer the target strain value is to 0.762. It is worth mentioning that this conclusion depends on the dataset. If the dataset is enriched with information from more sheet metal coils, especially if these new coils have significantly different hole expansion strain values at fracture than the current ones, the ideal point would likely change (as well as the ROC curve).

5 Conclusions and Main Advantages of the Proposed Solution

In this work, the performance of various ML algorithms was evaluated in order to predict edge cracking in sheet metal forming. Three stages were considered in this process: first, the generation of the dataset, second, the experimental setup of the models and, finally, the evaluation of the models. Based on the performance metrics carefully chosen for the edge cracking problem, the analysis of the results showed superiority of the majority voting ensemble models over the stacking ensemble or even the single classifiers.

It was concluded that the majority voting models achieve the best performance, with similar values for the different combinations tested. Among the single classifiers, the MLP and kNN algorithms perform the best, and are valid alternatives to the majority voting models, due to the reduced computational power required. The stacking models, however, show worse performance for low target strain values that the remaining models and, therefore, should not be considered for this type of application.

The analysis of the ROC curve obtained for the majority voting model confirmed the quality of the predictions and shows that the models perform better for target strain values close to the median of the strain values in the dataset. This leads us to conclude that the operational point of the ML models can play a significant role when they are put in practice.

In general, the capabilities of ML algorithms to predict the occurrence of edge cracking are satisfactory, and their application in an industrial environment is recommended. These applications range from the design phase of new components, to the identification of defect probability when using newly received materials, and contribute to the reduction of production costs and scrap rates.

Acknowledgements This research is sponsored by FEDER funds through the programs COMPETE – "Programa Operacional Factores de Competitividade" and Centro2020 – "Centro Portugal Regional Operational Programme", and by national funds through FCT – "Fundação para a Ciência e a Tecnologia", under the projects UIDB/00285/2020, UIDB/00326/2020, UIDB/00481/2020 and UIDP/00481/2020 and CENTRO-01-0145-FEDER-022083. It was also supported by projects PTDC/EME-EME/31243/2017 (POCI-01-0145-FEDER-031243), PTDC/EME-EME/31216/2017 (POCI-01-0145-FEDER-031216) and POCI-01-0247-FEDER-017762 (SAFEFORMING), co-funded by FCT, by FEDER through the program Portugal-2020 (PT2020), by POCI and by the Portuguese National Innovation Agency (ANI). The first author, Armando Marques, was supported by a grant for scientific research from the Portuguese Foundation for Science and Technology (ref. 2020.08449.BD). All supports are gratefully acknowledged.

References

Aghasafari, P., Abdi, H., & Salimi, M. (2014). Artificial neural network modeling of flow stress in hot rolling. *ISIJ International, 54*, 872–879.

Beskopylny, A., Lyapin, A., Anysz, H., Meskhi, B, Veremeenko, A., & Mozgovoy, A. (2020). Artificial neural networks in classification of steel grades based on non-destructive tests, *Materials, 13* (11), 2445.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing, 408*, 189–215.

Dib, M. A., Oliveira, N. J., Marques, A. E., Oliveira, M. C., Fernandes, J. V., Ribeiro, B. M., & Prates, P. A. (2020). Single and ensemble classifiers for defect prediction in sheet metal forming under variability. *Neural Computing and Applications, 32*, 12335–12349.

Fernández, D. M., Rodríguez-Prieto, A., & Camacho, A. M. (2020). Prediction of the bilinear stress-strain curve of aluminum alloys using artificial intelligence and big data. *Metals, 10*, 1–29.

Gorji, M. B., & Mohr, D. (2019). Towards neural network models for describing the large deformation behavior of sheet metal. *IOP Conference Series: Materials Science and Engineering, 651.*

Guevara, J. C., Torres, R. S., & Fonseca, N. L. S. (2020). On the classification of fog computing applications: A machine learning perspective. *Journal of Network and Computer Applications, 159.*

Hartmann, C., Opritescu, D., & Volk, W. (2019). An artificial neural network approach for tool path generation in incremental sheet metal free-forming. *Journal of Intelligent Manufacturing, 30*, 757–770.

Koenuma, K., Yamanaka, A., Watanabe, I., & Kuwabara, T. (2020). Estimation of texture-dependent stress-strain curve and r-value of aluminum alloy sheet using deep learning. *Materials Transactions, 61*, 2276–2283.

Konstantopoulos, G., Koumoulos, E. P., & Charitidis, C. A. (2020). Classification of mechanism of reinforcement in the fiber-matrix interface: Application of machine learning on nanoindentation data. *Materials and Design, 192*, 108705.

Masi, F., Stefanou, I., Vannucci, P., & Maffi-Berthier, V. (2021). Thermodynamics-based artificial neural networks for constitutive modeling. *Journal of the Mechanics and Physics of Solids, 147*, 104277.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Spathopoulos, S. C., & Stavroulakis, G. E. (2020). Springback prediction in sheet metal forming based on finite element analysis and artificial neural network approach. *Applied Mechanics, 1*, 97–110.

Yucalar, F., Ozcift, A., Borandag, E., & Kilinc, D. (2020). Multiple-classifiers in software quality engineering: combining predictors to improve software fault prediction ability. *Engineering Science and Technology, an International Journal, 23*, 938–950.

AI for the Tuna Fishing Industry Applications

Carlos Groba

Abstract Artificial Intelligence (AI) is used to help the tuna fishing industry to improve its day by day operations at sea. Tuna fishing vessels that fish with FADs (Fish Aggregating Devices) face an optimization problem in a dynamic scenario never seen before in other industries. Solving this issue can help this industry to minimize fuel consumption and emissions to the atmosphere. Considering the optimization challenge in greater detail, the problems to solve are two. The first is the basic case in which a tuna fishing vessel equipped with N buoys or FADs wants to know the best route to visit them all. The second goes further and tries to reach the same solution when a group of M vessels shares N FADs. In this second case, a more global solution is needed, including multiple vessels and more FADs to visit, but it can solve the global optimization problem for an entire fleet of tuna fishing vessels, with optimal results. The combination of AI algorithms and prediction is key to finding a solution for such a complex, dynamic environment. Specifically, a genetic algorithm (GA) is combined with a prediction method. This is an academic novelty and gives excellent results in comparison with the current industry standard and with the literature covering the best techniques for solving this problem. Moreover, the solution suggested is a new kind of global solution that is valid for both static and dynamic scenarios. The proposed solution is tested using real data from the fleet. Results show how profitable it is for the firm, as well as for the planet, greatly reducing emissions to the atmosphere and helping the tuna fishing industry to be more sustainable.

C. Groba (✉)
School of Economics and Business, S/N. Campus das Lagoas/Marcosende, Universidade de Vigo, Rúa Leonardo Da Vinci, 36310 Vigo, Spain
e-mail: cgroba@uvigo.es

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8_7

1 Introduction

1.1 The Tuna Fishing Industry

The tropical tuna fishery is one of the largest and most important in the world. Tuna fishing occurs in all major oceans of the world and has grown steadily over the past 60 years. The most widely used and fastest-growing fishing gear targeting tuna is the purse seine (Parker & Vázquez-Rowe, 2015).

In the open ocean, many species, including tunas, are associated with objects drifting on the surface such as logs or branches (Dempster & Taquet, 2004). Fishing around floating objects is associated with a higher successful haul, or 'set' rate than targeting free-swimming schools (Fonteneau, Pallares, & Pianet, 2000; Floch, Delgado de Molina, Assan, Dewals, Areso, & Chassot, 2012).

In the mid-1980s, skippers started experimenting with ways to maximize the potential of floating objects as fishing tools. Initially, reflectors and radio beacons were attached to logs to improve their detection over greater distances and fishers eventually started constructing purpose-built, drifting fish aggregating devices (FADs), fitted with electronic buoys to simultaneously boost the number of floating objects in the ocean and further aid their detection (Davies, Mees, & Milner-Gulland, 2014). The technology to track FADs continued evolving over the years, from radio-based buoys to GPS buoys that use satellite communication and are equipped with echo sounders to monitor the amount of biomass aggregated beneath the FAD. The advantage of FADs for increasing the catchability of tuna made the use of FADs skyrocket for the purse seiner industry, even for the fleets that traditionally had relied on free school (Lopez, Moreno, Sancristobal, & Murua, 2014).

From now on, we shall refer to either FADs or buoys to mean floating objects that drift in the open ocean for tuna fishing vessels to track and to fish beneath when possible.

FADs are made by fishermen who use simple materials like wood, string, and net, so two FADs cannot be assumed to drift identically under the same conditions. Figures 1 and 2 shows how a FAD is. In Fig. 1, we can observe the shape and materials used in the part that is floating on the water, meanwhile, Fig. 2 shows a typical composition of the FAD beneath the water. Although in Fig. 1 we see an industrial FAD, they are generally made of bamboo or pieces of wood and stabilized in the surface currents by large pieces of netting hanging below (Fig. 2).

1.2 Marine Instruments S.A. and Its Role

The industrial sector addressed in this chapter is the fishing industry, focused on tropical tuna fishing. The Marine Instruments S.A. company creates products and services for the maritime industry, mainly for tuna fishing, for which satellite buoys for tracking FADs (Fish Aggregating Devices) are their main product.

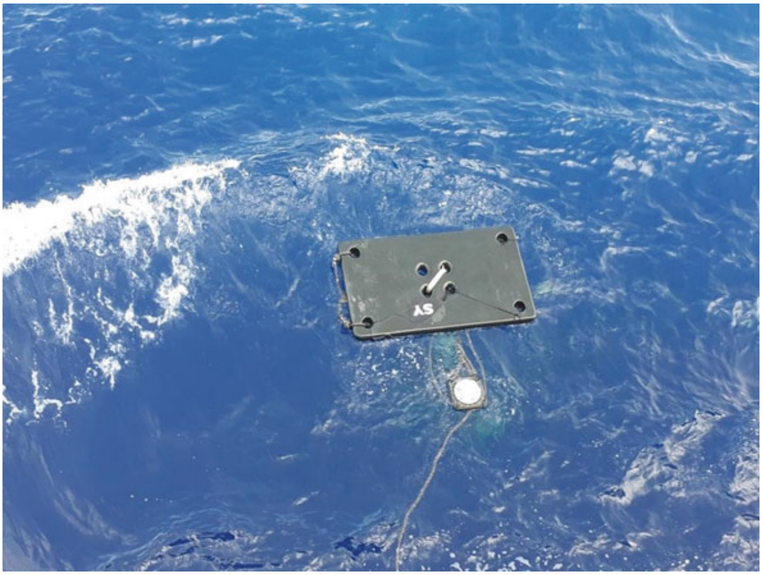


Fig. 1 FAD drifting in the ocean with a satellite buoy tied to it

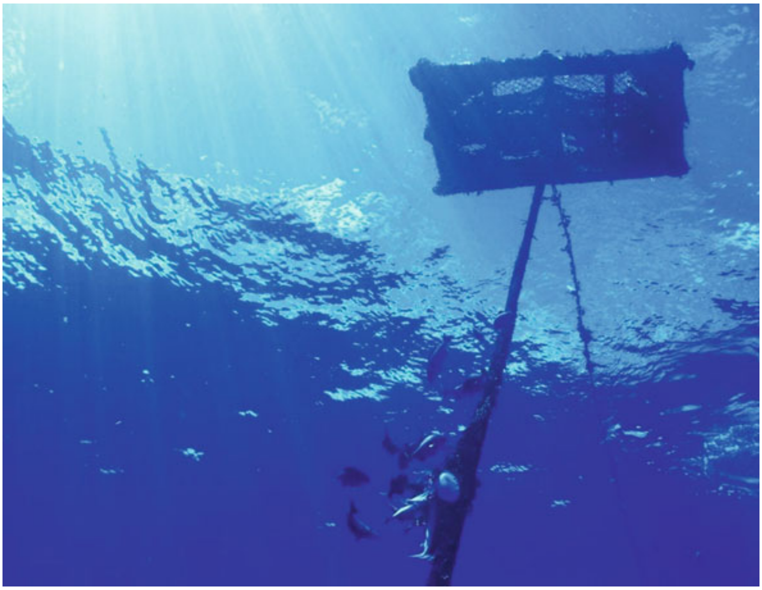


Fig. 2 FAD structure beneath the water

This Spanish company was created in 2003 and is located in Nigrán, a village in northwest Spain, close to the border with Portugal. Since its foundation, the company has seen steady growth, achieving a turnover of 40 M€ in 2019 with 130 employees, 60 of whom belong to the R&D department. With more than 40% of its staff working on R&D, it is clear that this company aims to improve its products and services while also creating new ones for different new markets. All its products are designed in the R&D department. Marine Instruments also has manufacturing capacity, based on the high-quality standards, it aims to guarantee for each of its products. Besides the fishing sector, Marine Instruments (MI) has solutions for the aquaculture industry and also for defense, among others.

Marine Instruments is one of the companies in the world that has taken part in the evolution of buoy technology for the tuna fishing industry. It began with the development of a GPS radio buoy for tuna fishing, moving from radio to satellite technology in 2005 (using Inmarsat D+), and from Inmarsat D+ to Iridium SBD in 2009. This year was also an important year because echo sounder technology was launched, marking a great improvement in the buoy industry and setting a new buoy standard. Buoys without an integrated echo sounder gradually disappeared from the tuna fishing industry.

The solution provided by Marine Instruments to the tuna fishing industry is not just the buoy itself, but a complete service covering the buoy, satellite communications between the buoy and servers (where all the information is stored), a satellite receiver for the vessels (MSR), and also specific software to receive all the information on board. This software, known as MSB, is able to connect to the MSR satellite receiver, download the buoy information, and plot it in a digital chart (Fig. 3).

The complexity behind the buoy market is shown in Fig. 3, where we can observe that the information gathered by the buoy is sent to a satellite constellation (typically

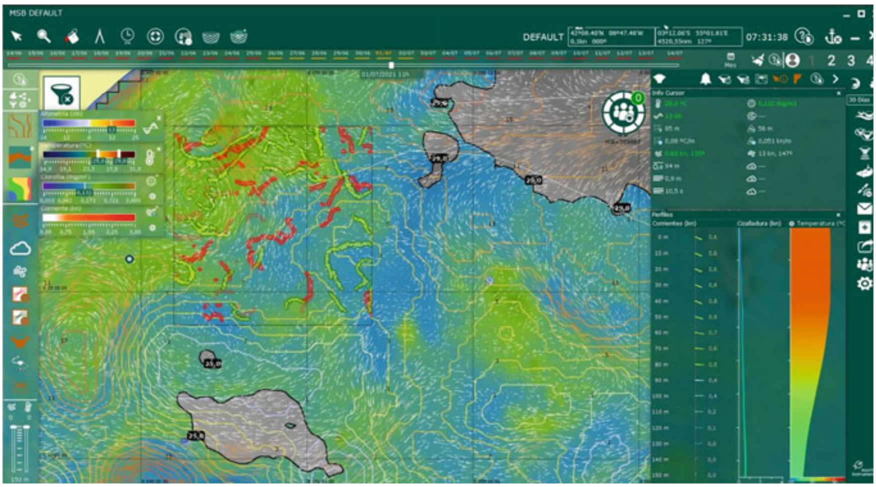


Fig. 3 MarineView software to receive buoys and observe oceanographic conditions

Iridium or Inmarsat) (Moreno, Dagorn, Sancho, & Itano, 2007). It is then sent from the satellite provider servers to the company's servers, where it is processed and sent on to the customers (vessels and offices) who can visualize it using specific software.

Thanks to this complete service, tuna fishing vessels can handle hundreds of buoys without difficulty, tracking their FADs, and taking the best decisions on where to fish.

2 Problem Statement

2.1 Simple Case

A tuna fishing vessel can be compared to a hunter, always looking for prey; in the case of tuna vessels, the prey can be free school tuna or tuna associated with FADs. During their search, they sail many miles, day and night.

Nowadays, a tuna fishing vessel may own around 300 buoys at the same time, but it can handle more if it shares its buoys with other vessels.

Every buoy is tied to a FAD, which is the platform that attracts tuna while it drifts in the ocean (Fig. 5). The buoy information travels through satellite communication, reaching the company servers and, after being processed, the information goes down to the vessel, again using satellite communication, as is represented in Fig. 3.

The reception of the information and its representation can be observed in both Figs. 3 and 4 (in detail), in which the software MSB/MarineView, besides the buoys information, also receives oceanographic information and allows the skipper to take

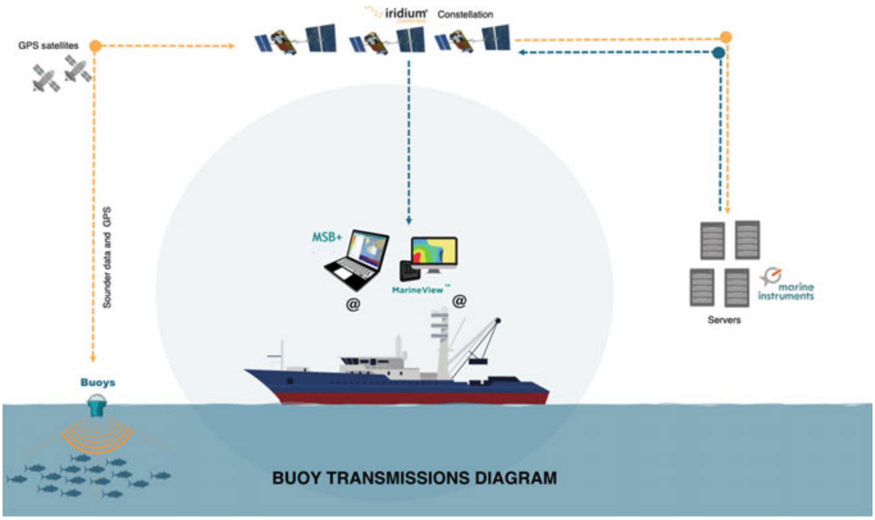


Fig. 4 Buoy—Server—Vessel schema

the best fishing decision very quickly. Nowadays, the information that a system like MarineView handles is:

- Buoys information (position, battery level, water temperature, and echo sounder)
- Currents (from 0 up to 150 m)
- Water temperature (from 0 up to 150 m)
- Temperature fronts
- Chlorophyll concentration
- Chlorophyll fronts
- Thermocline depth
- Shear currents
- Wind
- Sea surface height anomaly
- Clouds
- Wave height
- FAD drifting prediction
- Fish recommendations.

The software is prepared to be easy to use and to combine all the layers, thanks to a hardware console that allows the user to manage this information in an intuitive and easy way. With this, the user can combine different layers, use memories to save his most used configurations, and be quick in terms of taking the decision of where to fish.

Taking apart the oceanographic capabilities of MarineView. If we focus on the quantity of buoys that a single tuna vessel manages, what does a vessel have to do to fish? It certainly faces a giant problem of optimization every day. Each vessel has many options when choosing which buoys to visit and the best route.

As we can see in Fig. 5, a vessel faces an enormous range of possibilities at any one time. In fact, if a vessel handles N buoys, the number of different routes is N! This combinatorial problem is well-known as the Traveling Salesman Problem (TSP) (Gutin & Punnen, 2006). It has been studied for many years and there are several possible solutions.

Clearly, the problem faced by a tuna vessel is similar to the TSP, but if we look closer at the buoys and the vessel, we see that the real problem is different because the buoys move with the FADs as they drift with the different currents beneath them, so the problem changes with time, becoming even more difficult to solve (Fig. 6).

While the TSP is hard to solve, this new dynamic scenario seems to be even harder (Fig. 4). So how can we help the skipper to take the best decision and also to save fuel? The literature on this topic (Dynamic TSP or DTSP) is scarce, so this question represents the challenge presented in this study: how to find the best route in a dynamic environment.

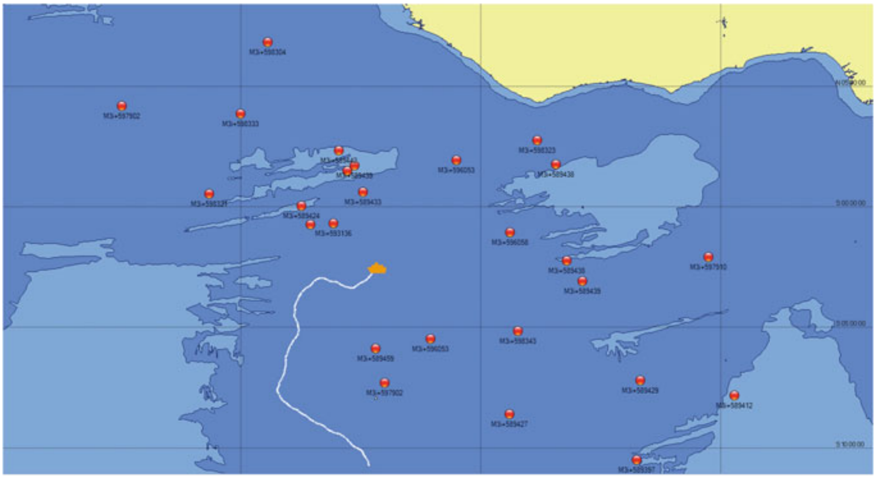


Fig. 5 Representation of a tuna vessel and its buoys in the Atlantic Ocean

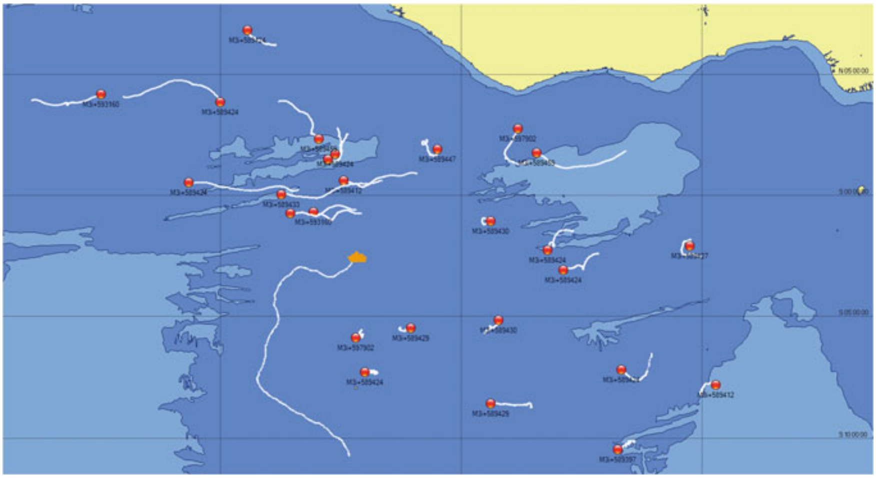


Fig. 6 A more realistic approach to the problem with dynamic buoys

2.2 Complex Case

As already stated, buoy technology evolved constantly over the years. From radio-based buoys covering only a few hundred nautical miles to satellite communication allowing for full coverage. Satellite communications imply the need for a server that concentrates all the information to be processed and sent to the vessel, which is very different from the point-to-point communication between buoy and vessel that radio-based buoys were based on. This intermediate server brought huge improvements: all the buoy information could be consulted by the manufacturer, so support improved considerably; secondly, tuna fishing offices could receive the buoy information and in some cases, an oceanographer could help vessels from the office; and thirdly, buoy information could be received by more than one vessel. The classic buoy-vessel route could be changed to buoy-server-vessels (Fig. 2), and this information could be changed at any time in the server, making changes for the fleet dynamic.

Satellite technology, therefore, eclipsed the former radio-based technology and, although each satellite buoy has a monthly cost in airtime, the changeover was worthwhile for fishing companies, which steadily increased their budget for satellite buoys, greatly improving their fishing results.

This new scenario in which buoys can be shared between vessels changed the rules of the game. Suddenly the bigger tuna companies started to see the advantages of working with two or three tuna vessels together as a group. In such cases, all the group's buoys were shared and, among other advantages, the logistics improved because they could distribute their vessels better in the ocean, unlike companies with a single vessel or those that continued working in the former fashion (not sharing buoys).

In this new scenario, the following question arises. When buoys are shared among vessels, what is the optimal route for visiting them? (Fig. 7).

This new challenge is obviously more complicated than working with only one vessel. The number of buoys is bigger, and the fact that there are multiple vessels raises the number of possible combinations, making the problem more difficult to solve.

This problem is known in the literature as the Multiple Traveling Salesman Problem (mTSP) or mTSP-MT (Moving Targets) for this specific case of a dynamic

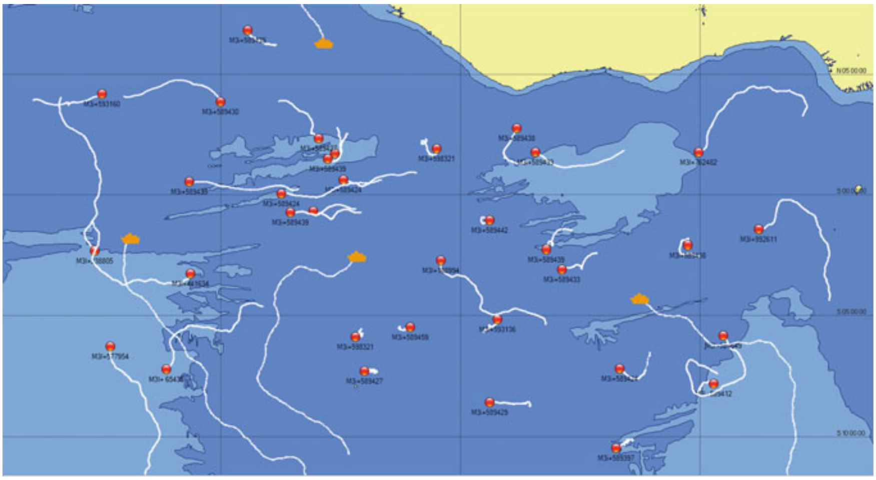


Fig. 7 Four tuna fishing vessels sharing buoys in the Atlantic Ocean

environment (Menezes et al., 2015). As with the dynamic TSP, the literature on mTSP-MT is still scarce (Groba et al., 2018).

So, to summarize, both scenarios require a solution for optimization. This is a new point of view for the literature, in this case, applied to the specific case of tuna fishing vessels. The first problem introduces the challenge to be solved and the second one goes further, adding greater complexity but allowing for a more global solution not only for a single vessel, but also for an entire fleet.

3 Solution Design

To search for a solution to this dynamic combinatorial problem, we considered three factors: the first was the prediction of the dynamic scenario caused by the moving buoys; the second was the use of Artificial Intelligence using evolutionary algorithms to solve the combinatorial problem of the optimal route, and the third and most important was a combination of the first and the second to achieve a new solution for the problem as a whole. With this combination, we get a new algorithm that evolves perfectly in this dynamic scenario and leads to a better solution to save time and fuel.

3.1 Prediction Method

In the dynamic scenario of drifting buoys tied to FADs, the proposed solution starts by trying, in this pseudo-chaotic environment, to figure out where the buoys are going to move in the near future.

FADs are man-made objects, generally consisting of bamboo rafts covered in old pieces of purse seine nets with several floats, but the most modern FADs are made of new materials to improve their buoyancy (Fig. 1). The subsurface structures found below FADs are typically made out of old fishing nets and extend to depths of 30–80 m. These subsurface structures create a sail effect for the ocean currents, but are also affected by waves and wind forces on the surface. Thus, the movement of FADs in the ocean depends not only on the meteorological and oceanographic conditions, but also on the materials they are made of, as well as the depth of the ropes. In addition, a FAD changes its drifting behavior with time because the structure on the surface and the ropes become heavier due to the different elements that attach to them (microparticles floating in the water, barnacles, etc.). Also, currents beneath FADs move in a range from 0.2 to 2 knots.

All these elements make it very difficult to accurately predict where a single FAD is going to move, the idea is not to know their exact location, but only where they are likely to be.

For this reason, the prediction method proposed is very simple. It does not involve complex models related to oceanographic currents or FAD structure modeling.

As FADs are tied to tracking buoys that send their position, we have information of the past movements of each FAD and, although we do not know how an individual FAD was made or the length of its ropes, we can see how it drifts every day. We also know, as stated above, that the FAD behaves like a big sail beneath the water, which means its movement has inertia. This was the clue for adopting a very simple method, namely, Newton's motion equation. This only needs the current and previous positions of the FAD to know its current speed and its acceleration. With only three positions, we can predict where the FAD will be in the future:

$$y_{t+1} = y_t + v_t \Delta_t + \frac{1}{2} a_t \Delta_t^2 \tag{1}$$

where:

y_t the position at the end of the interval (displacement)
v_t the velocity at the end of the interval t
Δ_t the time interval between the initial and current states
a_t acceleration at time t

Newton's motion equation was chosen because it offers a quick forecast of the future position of buoys or FADs from little information, without huge computational costs.

3.2 Artificial Intelligence Based on Genetic Algorithms

Genetic algorithms (GA) belong to the class of evolutionary algorithms (EA), which are inspired by natural selection. GA is a metaheuristic method that offers a sufficiently good solution to an optimization problem. Introduced by Holland (1992), GAs represent one of the most consolidated approaches to the TSP (Potvin, 1996).

Basically, GAs achieve the optimal solution from a random set of initial solutions called populations. Each set comprises an array of numbers where each number represents one of the targets on the route, which are named genes. Hence, each population is evaluated by a fitness measure (in our study, for instance, the measure is determined by the minimal distance between all points on each route), then parents of the next generation are selected probabilistically from the whole population, and the best routes are selected to become the parents of the next generation. The process is regulated by operators reflecting typical gene traits such as crossover and mutation. GAs repeat this loop until they converge to a near-global optimal solution (Groba et al., 2015). For instance, GAs can be used to solve a combinatorial problem such as the TSP, where the number of combinations makes it impossible to check all possibilities.

3.3 The Power of Combination

On the one hand, we have seen that with the help of a GA, we can solve a huge combinatorial problem like the TSP, which is similar to the problem that a tuna vessel faces when fishing with FADs. However, the dynamic nature of the ocean and the movement of FADs make this approach not valid at all. If we use the classic TSP approach to solve the tuna vessel combinatorial problem, the solution will consider that FADs do not move, so it is of no use for the vessel.

On the other hand, we used a very simple method to forecast the future positions of FADs. The aim is not to find the best forecasting method for such a dynamic environment as the ocean, but to prove that knowing where the FADs might be in the near future can help to find a better solution than the literature can offer for now.

With these two ideas in mind (TSP solution using GAs, and prediction), the key appears to be to combine forecasting with an evolutionary algorithm, in this case, a GA. If we can introduce the predicted position of FADs in the GA, the algorithm can evolve toward the dynamic situation of the FADs and, although the forecasting is far from perfect, it should offer a better global result than other methods.

4 Enhanced Genetic Algorithm

4.1 Simple Case

How can we integrate FAD forecasting into the GA design? The solution lies in the fitness of each chromosome. The TSP solution will be the route that the vessel must follow, and this route is formed by the FADs the vessel has to retrieve.

If we include forecasting, when a solution is created, the fitness calculation will indicate where the FAD is going to be, not just where it is. We know what the vessel speed is when recovering FADs, and how much time the vessel spends fishing at each FAD. With these two parameters, it is feasible to calculate the distance traveled by the vessel depending on its arrival time at each FAD, but calculating where the FAD is going to be at that time using Newton's equation.

With this combination, the GA will evaluate each solution in a similar way in this dynamic environment, thus evolving toward a better solution.

Figure 8 shows this concept graphically. The vessel represents the initial position of the vessel. The current position of each buoy is represented by a black symbol, and the previous track that the buoy traveled is shown by a continuous line with black dots. The predicted future movement of each buoy is marked by a dotted line. The white symbol shows the position where the vessel should be able to find the buoy in the near future.

We can observe how the algorithm calculates the fitness of a route. The vessel starts at time t and goes straight to the first buoy to recover it. The second buoy is recovered in a position to be indicated in the future. This is because the vessel spends

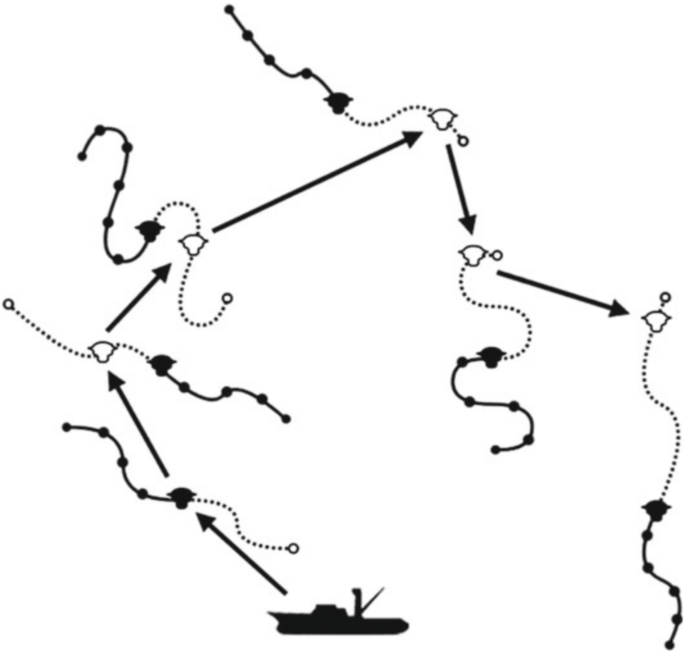


Fig. 8 Graphical representation of the algorithm evolution

some time traveling from its starting position to the first buoy, and we may have to consider fishing time at that buoy. This time spent by the vessel before moving on to the second buoy is used by the algorithm to predict where the second buoy is going to be at that time, which is more realistic than considering the buoy will not move at all. The same occurs for the rest of the buoys, using the time to travel and to fish at the previous buoys. For instance, to calculate where buoy N is going to be will depend on the time $t_1 + t_2 + \cdots + t_{N-1}$. To know where the vessel is going to be at any time t, we need to know the vessel speed and the time spent fishing at a buoy.

Therefore, to solve the problem, we need to predict where all FADs are going to be in the future. Then, based on their last positions and using Newton's motion equation, we can predict how each FAD is going to move in the future, as shown in Fig. 9. The matrix on the left shows the input information we need: for each FAD (f_i) we have the current position at t and the last two positions ($t - 1$ and $t - 2$). From the left matrix, using Newton's motion equation, we can calculate r, that is, the next future positions the algorithm needs for each FAD, represented in the matrix on the right.

The predicted positions are used by the algorithm to determine where each FAD is going to be at time t, depending on when the vessel is going to retrieve it, as Fig. 6 illustrates.

$$\begin{pmatrix} f_1^t & f_1^{t-1} & f_1^{t-2} \\ f_2^t & f_2^{t-1} & f_2^{t-2} \\ \vdots & \vdots & \vdots \\ f_n^t & f_n^{t-1} & f_n^{t-2} \end{pmatrix} \rightarrow \begin{pmatrix} \hat{f}_1^{t+1} & \hat{f}_1^{t+2} & \cdots & \hat{f}_1^{t+r} \\ \hat{f}_2^{t+1} & \hat{f}_2^{t+2} & \cdots & \hat{f}_2^{t+r} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{f}_n^{t+1} & \hat{f}_n^{t+2} & \cdots & \hat{f}_n^{t+r} \end{pmatrix}$$

Fig. 9 Prediction of FADs' future positions

Based on this prediction, the fitness value will change compared with the static case, and the algorithm will evolve toward a quasi-optimal solution, which suits this dynamic environment better than the static approach.

In Fig. 10, the block diagram of the algorithm is shown completely. First, on the left, we show the general algorithm, which follows the structure of most GAs except in the first step, the prediction of the future route of the objects (buoys) before starting, which is calculated as we have just explained.

Second, the main difference lies in how the fitness calculation is done. This is shown in the block diagram, on the right. In this case, the fitness uses the future position of each buoy based on the estimated arrival time of the vessel at that buoy. It changes the final route, so the algorithm evolves through a more realistic scenario and the results obtained are better than with other approximations, as we will see in the results.

4.2 Complex Case

A novel solution like the one presented here to solve the dynamic problem of tuna vessels when recovering FADs is promising, helping to save time and fuel, but, unfortunately, it is not enough in many cases. Vessels usually share FADs, generating a different optimization problem, which is bigger than the previous one.

Sharing FADs among vessels happens when they belong to the same company and operate in the same ocean. The number of vessels sharing information can range from two to five or more, depending on the fleet size. When FAD sharing takes place, all the FADs can be retrieved by all the vessels in the group.

Figures 5 and 11 shows the problem that arises when FADs are shared by three vessels. Even with the small quantity of buoys that we see in both figures, the complexity of the problem is huge, and it is clearly more difficult to solve than in the previous case.

If we consider that M vessels want to retrieve N buoys or FADs, then N is much larger than M ($N >> M$).

The previous solution, based on the well-known TSP, is insufficient. The literature proposes Multiple TSP, or mTSP, but for the problem of the dynamic nature of the elements to be retrieved, it proposes mTSP-MT (Moving Targets). This problem can

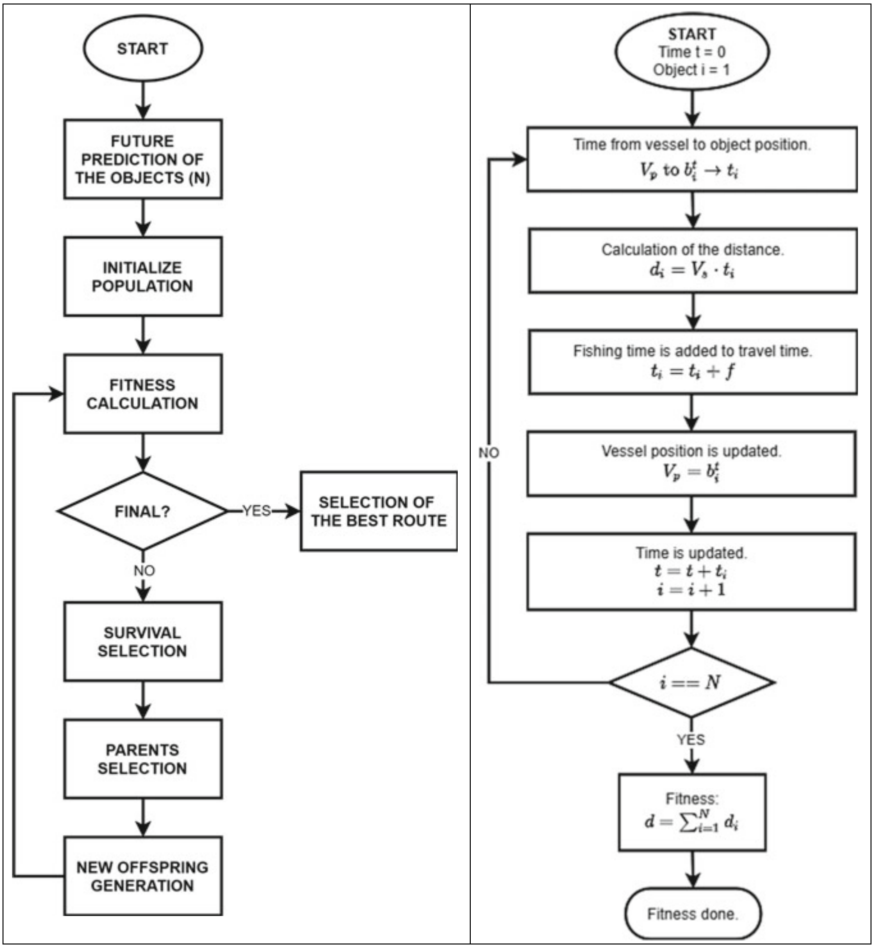


Fig. 10 General algorithm block diagram

also be solved from the perspective of the VRP (Vehicle Route Problem), which could be considered a generalization of mTSP with particular applications to transport and logistics (Montoya-Torres et al., 2015). The main difference between the classic moving TSP and the moving VRP is that the VRP can include additional restrictions apart from distance, such as added vehicle capacity, time constraint, a known non-negative demand for each depot, and a non-negative cost for each route (Eksioglu, Vural, & Reisman, 2009).

To solve the mTSP using the same GA approximation, each solution of the mTSP is a vector divided into two parts:

$$f_1, f_2, \ldots, f_N | z_1, z_2, \ldots, z_N) \tag{2}$$

Data: n FAD positions (f_i) at time, $t, t-1$, and $t-2$:

$$\begin{pmatrix} f_1^t & f_1^{t-1} & f_1^{t-2} \\ f_2^t & f_2^{t-1} & f_2^{t-2} \\ \vdots & \vdots & \vdots \\ f_n^t & f_n^{t-1} & f_n^{t-2} \end{pmatrix}$$

Data: m vessel positions: $(v_1, v_2, \ldots, v_m)$
Data: Vessel speed (s) and fishing time for each FAD
Result: Final route for each vessel
for $i \leftarrow 1$ ***to*** n **do**
  using input $(f_i^t, f_i^{t-1}, f_i^{t-2}) \rightarrow$ calculation of the r future positions of each target f_i: $(\hat{f}_i^{t+1}, \hat{f}_i^{t+2}, \ldots, \hat{f}_i^{t+r})$;
end
GA initialization: k first solutions calculated;
for $i \leftarrow 1$ ***to*** k **do**
  fitness(i)
end
while *Stopping criteria not reached* **do**
  parent selection;
  new offspring generation: k solutions calculated;
  for $i \leftarrow 1$ ***to*** k **do**
    fitness(i)
  end
end
final route = best fitness;

Fig. 11 Algorithm for solving the complex problem

The first part of the vector $(f_1, f_2, \ldots, f_N)$ represents all the targets (FADs) that must be visited, and the second part $(z_1, z_2, \ldots, z_N)$ shows how many targets each vessel must visit. Thus, N is the quantity of FADs to be recovered, and M is the number of vessels that need to pick up those N objects.

The generation of the initial population follows the method used by Carter and Ragsdale (2006). The first part of each chromosome is a randomly generated permutation of N FADs. The greedy solutions are subsequently generated by examining the present location of each vessel and then calculating the unassigned FAD that is closest to each vessel. Once a FAD is assigned to the closest vessel, the process continues until all FADs are assigned to a vessel, which gives the GA a good starting point in the search space and improves the results (Groba et al., 2018). The complete algorithm implementation can be observed in Fig. 9 as well as the fitness design in Fig. 11. On the other hand, Fig. 11 shows the complex problem graphically and Fig. 12 its solution (Figs. 13 and 14).

Data: Solution input: $(f_1, f_2, \ldots, f_n \mid z_1, z_2, \ldots, z_m)$
Data: FAD position matrix:

$$\begin{pmatrix} f_1^t & \hat{f}_1^{t+1} & \hat{f}_1^{t+2} & \cdots & \hat{f}_1^{t+r} \\ f_2^t & \hat{f}_2^{t+1} & \hat{f}_2^{t+2} & \cdots & \hat{f}_2^{t+r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{f}_n^t & \hat{f}_n^{t+1} & \hat{f}_n^{t+2} & \cdots & \hat{f}_n^{t+r} \end{pmatrix}$$

Data: Vessel position inputs: $(v_1, v_2, \ldots, v_m)$
Data: Vessel speed (s) and fishing time (the same for all vessels)
Result: fitness of the solution $\equiv$ total distance traveled
variable initialization: $z_0 = 1$;
for $i \leftarrow 1$ ***to*** m **do**
 $v = v_i$: initial position of the vessel i;
 $r = 0, d_i = 0$: time and distance equal to zero;
 $z_f = (z_0 + z_i) - 1$;
 for $j \leftarrow z_0$ ***to*** z_f **do**
  nt = time to recover $\hat{f}_j^{t+r}$ at s knots from position v;
  nd = distance traveled from position v to $\hat{f}_j^{t+r}$ at s knots;
  v = predicted position of $\hat{f}_j^{t+r}$: update the vessel's position;
  $r = r + nt+$ fishing time: update the time;
  $d_i = d_i + nd$: update the distance;
 end
 $z_0 = z_i + 1$;
end
fitness $= \sum_{i=1}^{m} d_i$

Fig. 12 Algorithm to calculate fitness for the complex problem

One of the most interesting options to investigate beyond the presented problems is to adapt the algorithm to quantity the fish that the buoys detect. In this case, depending on the echo sounding value reported, the algorithm has to balance the trade-off between the most logistic optimal route and the most pragmatic option that prioritizes the recovery of buoys with more fish beneath, even if they are far from the best route. It is important to note that if a buoy detects fish today, it does not mean that the fish will continue tomorrow, so maybe if the skipper visits that buoy in the next two days, the fish disappears. This is a practical problem that skippers deal every day, so our optimal route algorithm should do it too.

Nowadays, our research work follows this line. Figure 15 shows the problem and a possible solution for the specific case of three vessels sharing FADs. Sharing FADs means sharing buoys, for instance, the echo sounder information is shared too, and each buoy reports the quantity of fish detected per day. In Fig. 15, we can see the serial number of each buoy and the quantity of tons detected. The challenge is to modify the algorithm to balance the best combination between quantity with the optimal route.

For instance, the skipper may prefer to prioritize going first to the best buoys (more tons reported), but balance to not choose very bad route optimization, or maybe the opposite, the skipper could prefer to keep one of the best logistic routes losing some

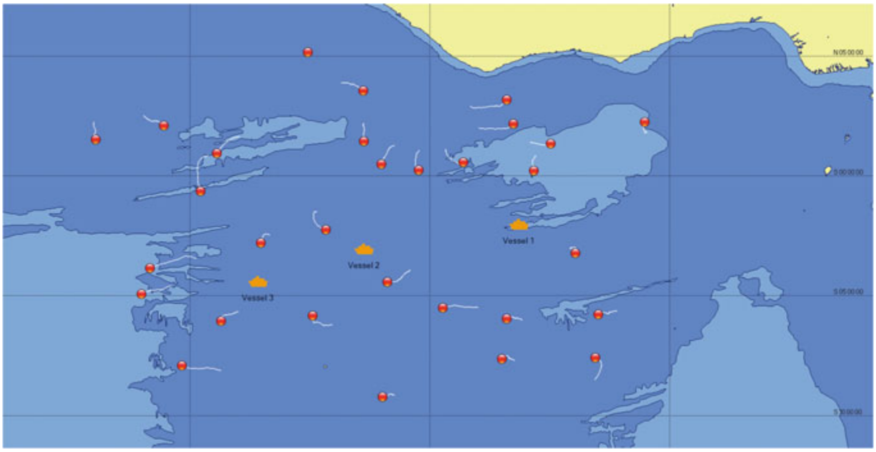


Fig. 13 Graphical representation of the problem in a complex case

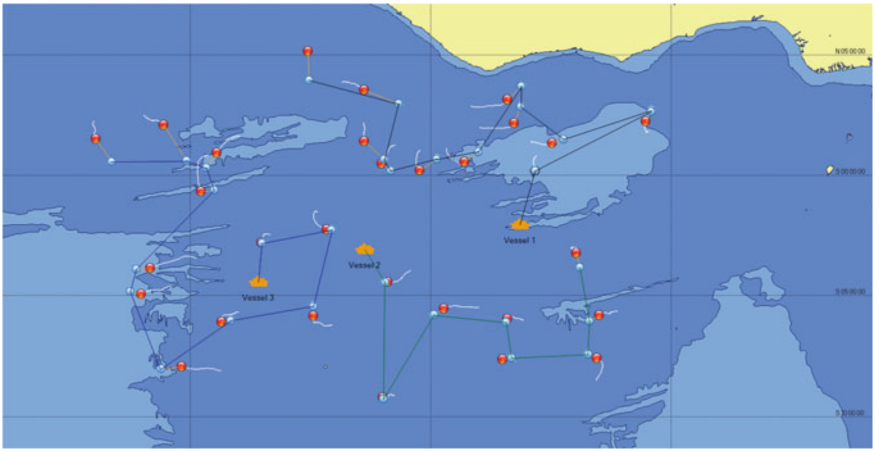


Fig. 14 Graphical representation of the solution in a complex case

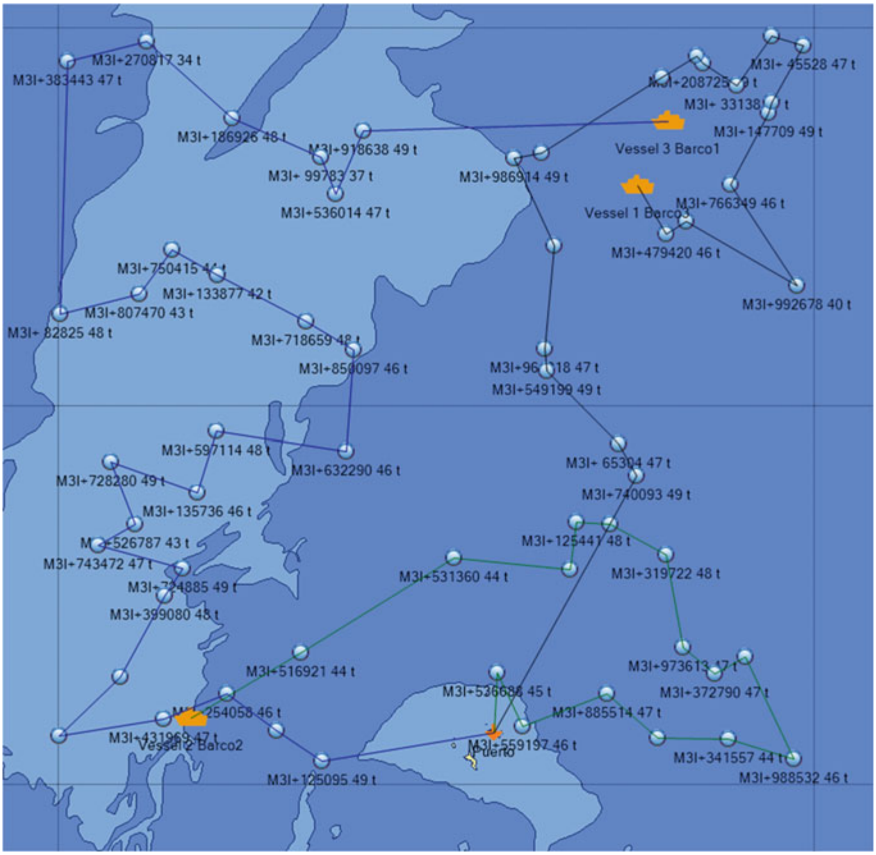


Fig. 15 Optimal route with three vessels, buoys, and tons information

miles because it considers the best buoys. Because of this, our findings show that a meta-parameter is necessary to tune this performance, that typically, the skipper can adjust to his preference. Currently, we are working in this line that makes our solution more practical for the skippers.

5 Computational Results

Both algorithms were tested with real data from buoys deployed in the ocean, provided by tuna fishing companies. To measure the performance achieved, both were compared with two different algorithms, the Nearest Neighbor (NN) approach and a genetic algorithm (GA). NN can be considered the simple method that skippers generally apply, and consists of always recovering the closest buoy to the vessel.

The second method used is a state-of-the-art GA without prediction.

Some assumptions were made to be able to run the algorithms properly. These assumptions include the vessel speed, as well as the recovery time for each FAD or the number of FADs to be recovered by each vessel.

Common assumptions.

- Vessel speed = 12 knots
- Recovery time = 3 h for each FAD
- FADs can take different speeds, ranging from 0.2 knots to 2 knots
- Distance between FADs is variable, ranging from 100 nm to 1.500 nm.

Assumptions for the simple case only.

- Number of vessels: 1
- Number of FADs to recover = 6, 9, and 12.

Assumptions for the complex case only.

- Number of vessels: 2, 3, and 4 (the typical range per group in practice)
- Number of FADs: from 20 to 36
- Number of FADs/vessel: from 5 (20 FADs for 4 vessels) to 14 (28 FADs for 2 vessels).

In order to test our model exclusively for scientific purposes, Marine Instruments provided us with anonymous real data from several tuna vessels fishing in the FAO capture zone no. 57 (Eastern Indian Ocean) from April 9th to April 23rd, 2017. According to internal company records, in this area, there were about 40 vessels operating at the same time. We performed 10 different measurements in each experiment, varying the positions of the FADs and the vessels, to obtain representative mean values for each case.

The performance of the algorithm shows better results for both the scenarios than other common optimizing strategies, showing that mixing prediction with genetic algorithms fits well for the FAD recovery problem. Therefore, we prove that integrating forecasting within a metaheuristic method, such as GA, can yield better results than the simple non-predictive version in a dynamic scenario.

For the simple case, results are shown in Fig. 16, where the algorithm is named GATP (Genetic Algorithm with Trajectory Prediction), and NN and GA-TSP are the other methods for comparison. The different algorithms were executed with real data for 6, 9, and 12 FADs. The results from our method are always better than the other two. Performance improves from 6.2 to 8.3% compared with NN, and from 1.2 to 5.5% compared with a simple GA.

For the complex case, the experiment is slightly different. We use different numbers of vessels (2, 3 and 4), repeating for each case the number of FADs to be retrieved (20, 24, 28, 32, and 36). In this experiment, our algorithm is named GAMTP (Genetic Algorithm based in Multiple Trajectory Prediction) and is compared with NN (Nearest Neighbor) and GA (a version of the GA without prediction). In Fig. 17, we can see on the *x*-axis the number of FADs to be retrieved. In this second experiment, the results are even clearer. The more complex the environment is, the more

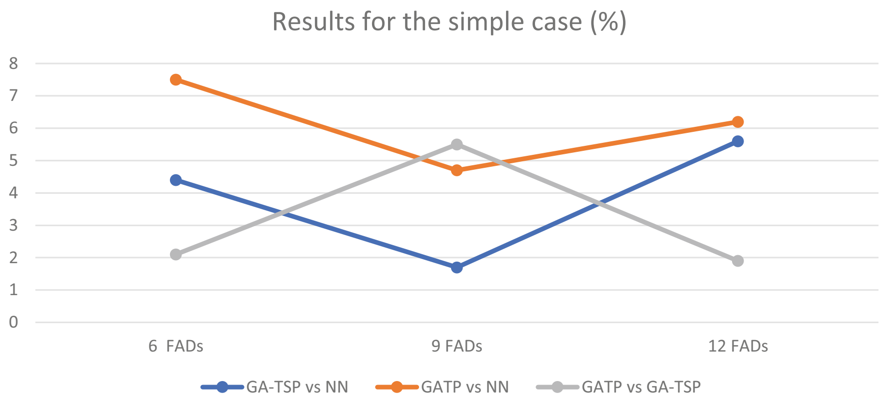


Fig. 16 Computational results for the simple case

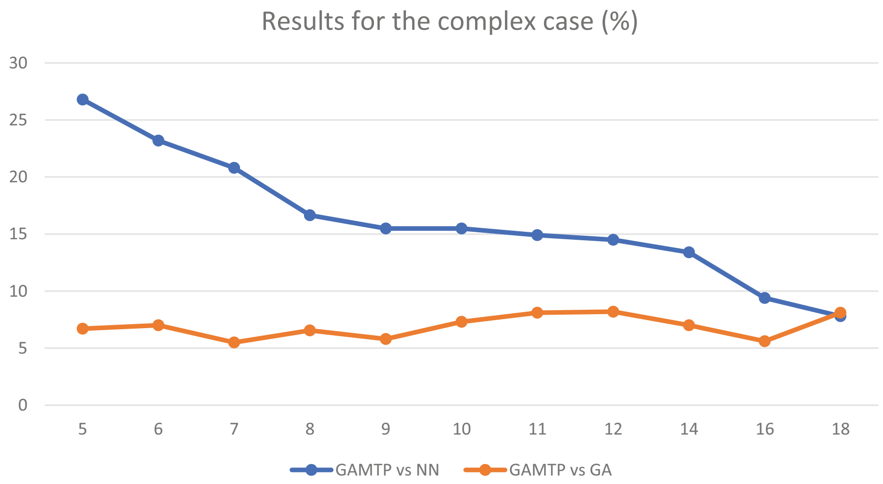


Fig. 17 Computational results for the complex case

options our solution offers to achieve improvements. For instance, with more vessels and FADs in the field, the results achieved are better than in the simple case.

The results of the complex case can be observed in Fig. 14, which shows the performance graphically. GAMTP gets better results in comparison with NN, rising from 12 to 25%. The improvement in relation to the GA without prediction is more stable, being constantly around 6.6%.

In spite of the general good results, in both the simple and complex experiments, we observe interesting behavior: the more FADs/vessel to recover, the worse our algorithm works. The explanation for such an interesting conclusion is simple, and the clue lies in the prediction algorithm. The prediction method used is very simple and does not work well when it predicts FAD movement for 3 or more days. Assuming that the complexity of FADs moving in the ocean is very high and that it depends on many factors and considering that we are simplifying matters a lot by using a prediction method like Newton's motion equation, then the more FADs to recover, the more time is needed by the vessels to retrieve them, so the predicted position of the last ones will be far from their current positions. This worsens the performance of our algorithm. When the prediction is not accurate, the algorithm converges to a solution in which the expectation of positions is wrong.

The conclusion is easy. It is better to not predict than to predict inaccurately. However, it is encouraging to know that, if we can improve the prediction method, the performance of the algorithm might improve even more.

The results are statistically supported and were tested using a Repeated Measures ANOVA. They were consistent due to the high significance of the most frequent multivariate test used.

6 Conclusions

This work shows the algorithms and techniques applied to improve efficiency in dynamic environments, such as that of tuna fishing vessels. The solutions presented, based on AI, focus on the tropical tuna fishing industry with FADs. This is a complex scenario that is appropriate for such techniques, although the algorithms developed could also be used for more general purposes.

In this context, the first study presents a novel solution with respect to the state of the art. Prediction and heuristic techniques are combined to achieve solutions that give a clear advantage as regards efficiency.

The second study focuses on solving a similar but even more global and more complex problem, that is, when several vessels share FADs.

The proposed algorithms not only solve the problems presented, but also allow for a more extensive solution, valid for both static and dynamic environments. All presented solutions were validated with simulations using real data.

The results show that the proposed solution works for both cases, improving by far on results using other optimization techniques. The algorithm could be improved with complex scenarios, but the easy prediction technique used here limits such an improvement due to its long-term error. This limitation could be minimized with more complex prediction methods.

Apart from the general theoretical results of these works, the algorithms presented are useful for the tuna fishing industry in the field, from which we received the necessary data for the simulations and for which important improvements in terms of efficiency were achieved. Due to the general nature of the algorithms and the solutions developed, it is reasonable to assume that the results will also be valid and applicable in other sectors where a dynamic logistics optimization problem exists. In addition to the direct savings achieved in distance and time, this work has a direct application for sustainability, because it helps to considerably reduce emissions to the atmosphere.

The algorithm presented is a general solution that is valid not only for dynamic scenarios, but also for static ones. These can also be solved easily when prediction is equal to zero.

With the idea of showing the tuna fishing industry the benefits of sharing FADs (such as the route optimization that the fleet as a whole can achieve), interesting work was carried out after this research, using a game-theory perspective and showing that, with the correct incentives, all stakeholders, including the company, the skipper, and even the environment can achieve mutually improved results by sharing FAD information (Groba, Sartal, & Bergantiño, 2020). This research opens up new and bold opportunities for companies because it helps them to share FADs among the entire fleet and to apply proper route optimization.

This research also opens up interesting research avenues aiming to improve the algorithms, adapting them to the real conditions in which tuna vessels work, including operating restrictions, i.e., hours when vessels do not fish, they need to return to the harbor, and other limitations beyond the initial theoretical approach.

References

Carter, A. E., & Ragsdale, C. T. (2006). A new approach to solving the multiple traveling salesperson problem using genetic algorithms. *European Journal of Operational Research, 175*(1), 246–257.

Dempster, T., & Taquet, M. (2004). Fish aggregation device (FAD) research: Gaps in current knowledge and future directions for ecological studies. *Reviews in Fish Biology and Fisheries, 14*(1), 21–42.

Davies, T. K., Mees, C. C., & Milner-Gulland, E. J. (2014). The past, present and future use of drifting fish aggregating devices (FADs) in the Indian Ocean. *Marine Policy, 45*, 163–170.

Eksioglu, B., Vural, A. V., & Reisman, A. (2009). The vehicle routing problem: A taxonomic review. *Computers & Industrial Engineering, 57*(4), 1472–1483.

Fonteneau, A., Pallares, P., & Pianet, R. (2000). A worldwide review of purse seine fisheries on FADs. In *Pêche thonière et dispositifs de concentration de poissons, Caribbean-Martinique, 15–19 Oct 1999.*

Floch, L., Delgado de Molina, A., Assan, C., Dewals, P., Areso, J. J., & Chassot, E. (2012). *Statistics of the European purse seine fishing fleet and associated flags targeting tropical tunas in the Indian Ocean (1981–2011).*

Groba, C., Sartal, A., & Vázquez, X. H. (2015). Solving the dynamic traveling salesman problem using a genetic algorithm with trajectory prediction: An application to fish aggregating devices. *Computers & Operations Research, 56*, 22–32.

Groba, C., Sartal, A., & Vázquez, X. H. (2018). Integrating forecasting in metaheuristic methods to solve dynamic routing problems: Evidence from the logistic processes of tuna vessels. *Engineering Applications of Artificial Intelligence, 76*, 55–66.

Groba, C., Sartal, A., & Bergantiño, G. (2020). Optimization of tuna fishing logistic routes through information sharing policies: A game theory-based approach. *Marine Policy*, *113*, 103795.

Gutin, G., & Punnen, A. P. (Eds.). (2006). *The traveling salesman problem and its variations* (Vol. 12). Springer Science & Business Media.

Holland, J. H. (1992). *Genetic Algorithms. Scientific American, 267*(1), 66–73.

Lopez, J., Moreno, G., Sancristobal, I., & Murua, J. (2014). Evolution and current state of the technology of echo-sounder buoys used by Spanish tropical tuna purse seiners in the Atlantic, Indian and Pacific Oceans. *Fisheries Research, 155*, 127–137.

Moreno, G., Dagorn, L., Sancho, G., & Itano, D. (2007). Fish behaviour from fishers' knowledge: The case study of tropical tuna around drifting fish aggregating devices (DFADs). *Canadian Journal of Fisheries and Aquatic Sciences, 64*(11), 1517–1528.

Montoya-Torres, J. R., Franco, J. L., Isaza, S. N., Jiménez, H. F., & Herazo-Padilla, N. (2015). A literature review on the vehicle routing problem with multiple depots. *Computers & Industrial Engineering, 79*, 115–129.

Menezes, M. B., Ketzenberg, M., Oliva, R., & Metters, R. (2015). Service delivery to moving demand points using mobile servers. *International Journal of Production Economics, 168*, 158–166.

Parker, R. W., Vázquez-Rowe, I., & Tyedmers, P. H. (2015). Fuel performance and carbon footprint of the global purse seine tuna fleet. *Journal of Cleaner Production, 103*, 517–524.

Potvin, J. Y. (1996). Genetic algorithms for the traveling salesman problem. *Annals of Operations Research, 63*(3), 337–370.

Application of Machine Learning Techniques to Automatic Energy Incident Detection

Rubén A. Gayoso Taboada and Javier Roca Pardiñas

Abstract Heating and cooling account for approximately half of the EU's energy demand. Nowadays, new heating, ventilation, and air-conditioning machines have sensors that generate huge amount of data. The analysis and processing of data from hundreds of devices and field sensors using Machine Learning (ML) models allows to address maintenance management. This enables the anticipation of failures and the scheduling of predictive maintenance. These lead to the optimization of processes, the minimization of downtime and costs, and more efficient use of energy. With the use of an ML model capable of predicting and organizing a wider range of incidents, an energy saving of 7% per installation is estimated, with a reduction of 30% in preventive maintenance visits and a reduction of 20% in corrective maintenance. More than 20% of the incidents registered could be corrected remotely without the need for a technician. Predictive maintenance will prolong the life of customers' HVAC machines and reduce the number that needs to be replaced.

1 Introduction

1.1 Brief Description of EcoMT and Its Activities

Ecomanagement Technology S.L. (EcoMT) (https://ecomt.net) is an IT company headquartered in A Coruña (Spain) that develops and integrates remote control and monitoring of facilities. The company belongs to a holding group formed by more than 25 companies, which has been undertaking engineering projects, conditioning, maintenance, and construction since 1983.

EcoMT has a multidisciplinary team of more than 50 employees, made up of telecommunications, industrial and computer engineers, mathematicians, and top

R. A. Gayoso Taboada (✉)
Instituto Tecnológico de Matemática Industrial (ITMATI), Santiago, Spain

J. R. Pardiñas
Departamento de Estadística e I.O, TMATI, Universidade de Vigo, Vigo, Spain

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8_8

technicians in industrial cooling and in electrotechnical installations. The company has been operative since 2010 and designs solutions in control, implementation of software, generation of algorithms to translate data into information, monitoring and tracking installations, renewable energies, and energy efficiency projects.

EcoMT bases its business model on its OTEA product, a remote control and monitoring software that allows its customers to be more efficient in managing the energy consumed by their facilities as well as the resources and supplies they need.

OTEA is an integrated system (hardware, software, services) for supervision and remote control of installations of HVAC, lighting, multisite, energy efficiency, and efficient maintenance plus enhanced comfort. OTEA is an experienced, reliable, and safe solution that has been available for almost 10 years, compatible with many types of hardware and machines on the market: open system.

Currently, OTEA (see Fig. 1) can monitor, analyze, and manage the behavior of air-conditioning and lighting equipment in more than 3,800 installations, controlling more than 50,000 machines mostly from the commercial segment. It is established in more than 50 countries all over the world, installed in more than 600 cities, and has been translated into eight languages. Each installation manages in real time more than 100 signals mainly related to temperature, power, and the state of machines. This totals more than 600,000 variables, resulting in 3 trillion records during the last 5 years. New facilities, with a growing number of machines producing huge amounts of data, require the necessary capacity of storage and processing, where Machine Learning (ML) models and statistical techniques may be applied.

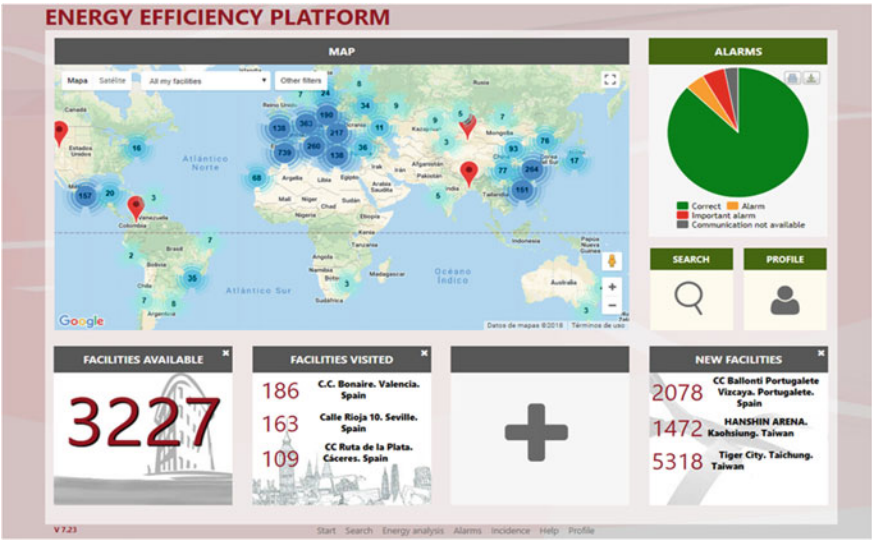


Fig. 1 Main screen of OTEA

1.2 Technology Transfer from University to Industry

By carrying out research, Universities are the main generators of new scientific knowledge. In the case of Spain, this is especially true of public universities. In order for this knowledge to be used by the whole of society, it is necessary to achieve effective knowledge transfer between universities and companies. In this way, the development of new technologies will go hand in hand with their implantation in the industry, thus providing new and innovative products and services.

The use and application of Machine Learning by companies is a clear example of a successful case of knowledge transfer from the university to the company. University researchers are working on the development of new algorithms to improve the classification of more and more data and to make increasingly more reliable predictions. However, the transfer and collaboration between science and business in Spain is not as effective as would be desired and compared to other European countries.

With regard to mathematics, researchers from the public universities of the Autonomous Community of Galicia are a national and international benchmark. In addition, in the field of transferring mathematical technology to industry, in 2012, its projects represented about half of all activity in contracts and projects with companies nationwide.

To consolidate the relevant position in the national and international field of industrial mathematics in Galicia, the University of A Coruña, the University of Santiago de Compostela, and the University of Vigo constituted, in 2013, the Consortium Technological Institute for Industrial Mathematics (ITMATI) (http://www.itmati.com), with the aim of becoming a technological center of international reference in the field of industrial mathematics.

ITMATI represents an important milestone in the aggregation of resources between the three universities to promote the transfer of mathematical technology and provide efficient and agile responses to the demands of companies, industries, and public administrations. Its main mission is to contribute to the strengthening and enhancement of competitiveness in the industrial and business environment, support innovation in the productive sector by achieving excellence in research, and the development of advanced mathematical technology oriented toward transfer to industry.

ITMATI has its own legal entity, which allows it to be totally autonomous when executing all types of activities. It streamlines and facilitates all the administrative procedures for the contracting of its services by companies and administrations. It groups together, in a single portfolio of technologies, the capacities and knowledge of researchers from the three universities and it presents all the potential of mathematics as a practical technology with great impact on a wide range of processes and businesses.

This chapter describes a successful example of the work of ITMATI in the transfer of mathematical technology, the case of Machine Learning to the company.

2 The Challenge

2.1 Heating and Cooling Energy Demand

Heating and cooling homes and industries account for half of the EU's energy demand European Commission (2021a). Moreover, 84% of the energy needed for heating and cooling is still generated from fossil fuels, while only 16% comes from renewable energy sources. In order to fulfill the EU's climate and energy goals, the heating and cooling sector must sharply reduce its energy consumption and cut its use of fossil fuels. In February 2016, the Commission proposed a common heating and cooling strategy for the EU. This is a first step in exploring the issues and challenges in this sector and solving them with common energy policies.

Installing efficient equipment and implementing efficiency improvement plans is no longer just an option, it is becoming an obligation. The EU has set a 20% energy savings target by 2020 (European Commission, 2021b). On 30th November 2016, the commission proposed an update to the energy efficiency directive including a new 30% energy efficiency target for 2030, and measures to update the Directive in order to make sure the new target is met. In addition, it is intended to save on the costs of each installation, both for a lower need of dedicated staff and for fewer repairs in machines that will be achieved by providing predictive maintenance.

The aim of early detection of incidents through an expert remote management system is an optimal use of energy and an important reduction in maintenance costs. This approach must, on the one hand, minimize the risk of unexpected failures which may occur before the next scheduled maintenance visit and, on the other hand, reduce the amount of unnecessary maintenance activities. Consider, for example, the case of a machine with dirty filters, which will need more energy to function. In this case, the developed models will detect the anomalous performance of the machine, allowing to correct it as soon as possible, hence favoring energy savings. This would also imply greater client satisfaction.

2.2 The Software OTEA

OTEA is a platform based on software for the monitoring and control of equipment, the main purpose of which is telemanagement and corrective maintenance of facilities. Developed by EcoMT, its design philosophy is to adapt to the needs of users in a modular way. OTEA allows the control and automation of installations using local hardware (PLCs and HMIs) and the centralization of existing PLCs and SCADAs.

Monitoring energy equipment such as heating, ventilation, and air conditioning (HVAC) implies handling huge amounts of data. All this information needs to be properly processed. It is, therefore, necessary to apply Machine Learning (ML) coupled with mathematical and statistical techniques. The use of High-Performance Computing (HPC) is also needed. In this case, the HPC and Big Data services were provided by the Galician Supercomputing Center (CESGA) (https://www.cesga.es) and the company Gompute (https://www.gompute.com).

HPC infrastructures offer the required capabilities to support these services and to provide scalable resources when necessary. Because of these needs, it is conceived as a new system with an aim to make the leap to the new industrial revolution, seeking to create eco-efficient intelligence by anticipating incidents related to demand and comfort.

With the use of these techniques, it was possible to deploy a predictive tool capable of detecting maintenance incidents before they happen, based on historical and real-time data and monitored by means of the platform of remote management of facilities, OTEA. The implemented tool applies mathematical and statistical algorithms in order to achieve reliable predictive maintenance, improving the efficiency of the current system, extending the equipment life cycle, and allowing a "smarter" resource management.

Before the use of Machine Learning (ML) coupled with mathematical and statistical techniques, a deterministic system with brute force conditions was applied to detect comfort anomalies through the OTEA platform, without considering previous problems in a personalized way (see Fig. 2). Such anomalies may be differences from setpoints above a given value or events out of schedule, to name two. This detection model has the consequence of obtaining false positives.

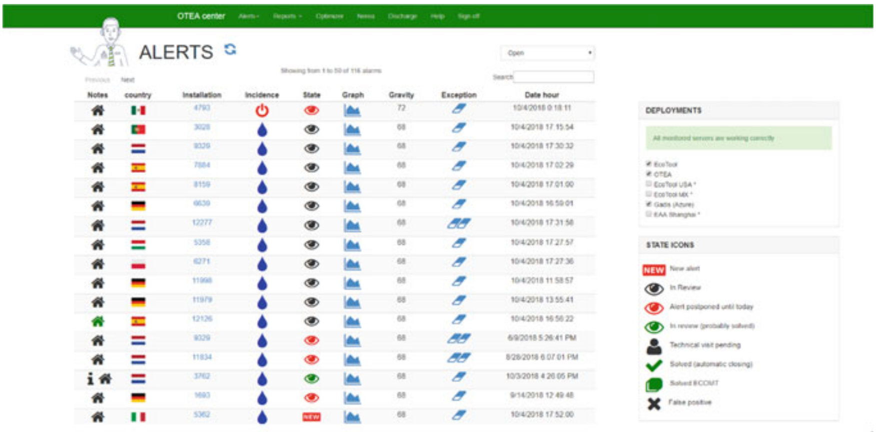


Fig. 2 Detection of anomalies by means of a deterministic system

3 The Solution

3.1 Predictive Maintenance

EcoMT's intention was to go further and address management in an innovative way, by using advanced ML techniques to create coherent mathematical models through the analysis of data collected from the hundreds of devices and field sensors located in the facilities. Among some of the challenges associated with the Industry 4.0 paradigm, EcoMT focuses on data generated by sensor networks from machines, and in its adequate processing. This valuable information enables the anticipation of failures and the scheduling of predictive maintenance that lead to the optimization of processes, thus minimizing downtimes and costs and improving energy efficiency. There is an ever-growing amount of detailed data that makes it necessary to apply ML on an HPC platform in order to make early decisions and to avoid breakdowns, based on historical and environmental data.

It is estimated from the samples used for validation, that the simple fact of forgetting to switch the operating mode from manual to automatic during the nights, which is very often the case, can increase electricity demand due to climatic conditions by 7%. With this development, facilities are continuously monitored so that early detection of anomalies can avoid unnecessary consumption. The translation of this figure into currently managed facilities would mean a great amount of financial savings yearly, which could be multiplied by a factor of 10 if all the facilities of these customers were controlled.

Once ML models and statistical algorithms were selected, the validation was performed by applying them in real and current scenarios provided by EcoMT, where it was found that the response is fast and accurate.

Predictive maintenance (PdM) evaluates the state of the machinery and recommends intervening or not, which produces great savings. The objective of this type of maintenance is to optimize the reliability and availability of machinery and critical equipment at a minimum cost (Fig. 3). Monitoring machines is nothing new but, in the context of Industry 4.0, the development of this activity allows incorporating in the machine sensor-generated data. The large number of records that need to be processed makes it necessary to use HPC.

This type of maintenance is understood as a set of instrumented techniques to measure and analyze variables to characterize the potential failure modes of productive equipment. To this end, an intelligent module was developed in OTEA that allows end users to make decisions in advance.

For the study development, data blocks belonging to 3,000 installations located throughout the world were used.

This model consists of four cyclic states, shown in a personalized way in Fig. 4, which are developed in the following order:

A. **Detection and processing of the signal**: the information is collected and pre-processed so as to introduce the data in a suitable format.

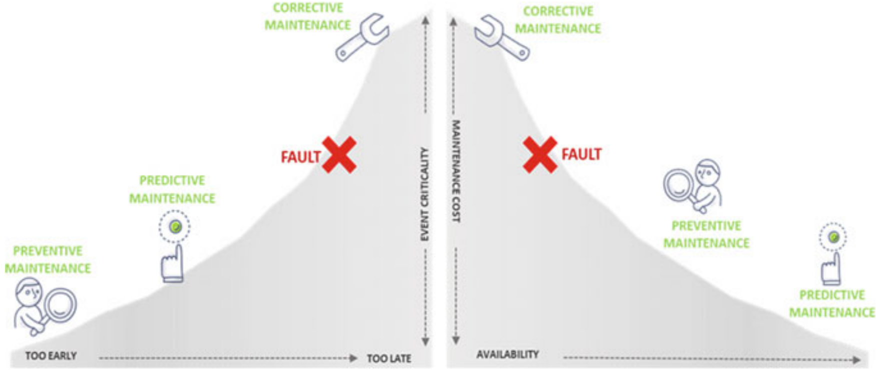


Fig. 3 Optimal maintenance timing: higher availability and lower cost

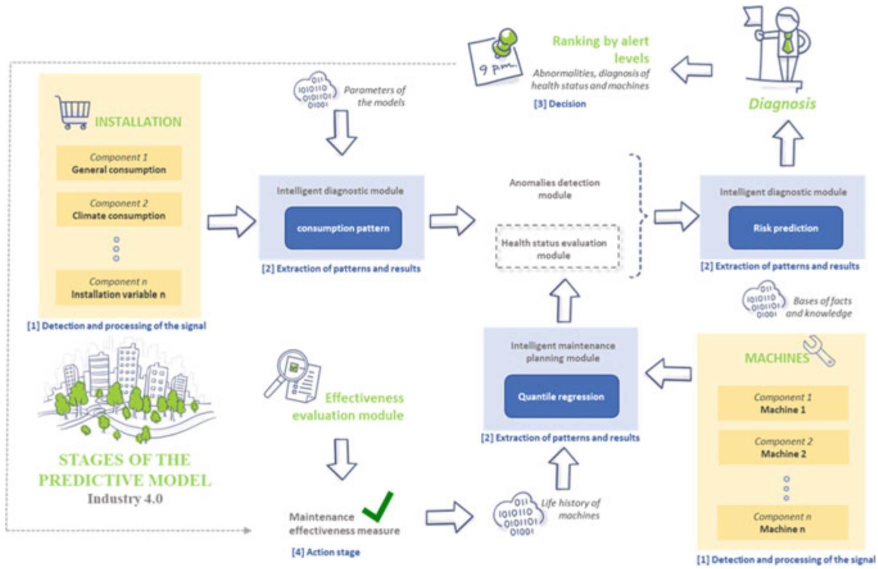


Fig. 4 OTEAres proposed design: predictive model in Industry 4.0

B. **Extraction of patterns and results**: with the information processed, the intelligent patterns and answers are extracted, applying ML models that were validated by using another block of current data and by comparing detected anomalies with those registered.

For EcoMT, it was interesting to analyze the behavior of each installation on two levels:

- Global: behaviors of climate demand and general demand are analyzed. These variables define the installation globally.

- Individual: each facility can have one or more climate machines, so the behaviors of each machine are analyzed. These variables define the individual performance of each piece of climate equipment, allowing to identify which are working correctly and which are not.

Once the two scenarios have been analyzed, a risk prediction is made where the studies are taken into account at both levels.

C. **Decision**: the contexts are processed and the actions to follow are established, with a diurnal and a nocturnal scenario. Once all the facilities have been analyzed, considering both the demand behavior and the operation of the climate machines, an individual diagnosis is made of each of the facilities and the tool provides a classification from higher to lower alert level (Fig. 5). The diagnosis to classify each installation in any of the six considered alert levels (from level 1 to level 6) is concluded from the behavior of the different models applied. In particular, it is verified if the model follows the pattern and if the machines have the correct performance, also taking into account the risk percentage. The latter is classified into five risks, which predict the expected dangerousness of the state of the facility.

D. **Action stage**: in this last step, the plan to be followed is chosen. In this case, this ranking will serve as support for the Control Center staff to improve management of the facilities, since it will act on those that show a higher incidence risk. Thanks to this hierarchical order, the incidences of unusual consumption and lack of comfort are reduced.

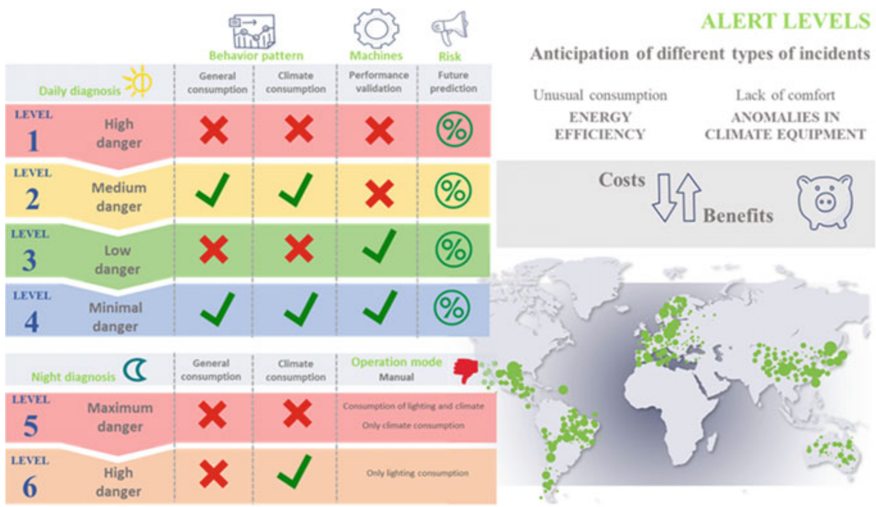


Fig. 5 Alert levels matrix

3.2 Development of ML Algorithms

EcoMT clients have several Programmable Logic Controllers (PLCs) installed in their premises which collect data about heating, ventilation, and air-conditioning (HVAC) systems. These systems transfer data periodically to a relational database at EcoMT. The collected data is used by the OTEA platform to monitor the systems and to release warnings and alarms when failures occur.

The main technical objective was to deploy a predictive tool that enables EcoMT to detect maintenance incidents before they happen, for which purpose Big Data and ML tools are used. Taking the existing data source as the starting point, several requirements and tools were identified. The new modules will fulfill several requirements in order to allow the development of a predictive tool based on ML. The first step was to determine the relevant metrics. This was done taking into account EcoMT experience and the availability of these metrics for most of the facilities.

It is important to note that each facility can have multiple machines and some of these metrics correspond to each machine. Hence, the number of metrics available for each facility may vary. In addition, the data was filtered to eliminate abnormal values using boundaries provided by EcoMT.

3.2.1 Detection of Active Power Patterns

The first step was to detect the active power pattern followed by each installation. The data registered at OTEA include the instant active power measured every few minutes. However, it was observed that these values could vary a lot when they are measured at this frequency. Hence, it was decided to aggregate the data by hours. The variables used to obtain the patterns were related to every climate machine. There were also selected features for training the models in several ranges of hours and times of the day as a categorical feature.

In order to be able to study the risk of incidence in a certain store and to be able to predict this phenomenon one day in advance, it has been necessary to construct a new variable, "incidence", from the data set and the information provided by the business. To obtain this new variable, we have considered a new setpoint temperature initially proposed by the company. This new variable takes into account the operating regime of the machine, either in cooling or heating mode. The new setpoint temperature (TC) is set as:

TC = 22 °C, if $TI \leq TA$(cooling regime).
TC = 20 °C, if $TI > TA$ (heating regime).

where TC denotes the setpoint temperature, and TA is the airflow temperature. Using this new setpoint temperature, we have calculated the binary variable "incidence", which will take a value of 1 when the ambient temperature differs from the setpoint temperature by ±2 °C in periods of time greater than 2 h (presence of incidence) and 0 otherwise (no incident). However, it has been observed that this variable does not obtain good results for predicting the risk of incidence, since we are grouping both the incidents that occur due to both an increase and a decrease in temperature. This can be seen in the left panel of Fig. 6, where a box plot is represented for the mean daily ambient temperature, taking into account the presence/absence of incidence (daily). Clearly, the temperature distribution is very similar both in the presence and in the absence of the phenomenon.

This situation led us to create two new incident variables (instead of just one), in order to be able to differentiate between the incidents caused during the cooling regime ("Inci1") and during the heating regime ("Inci2") (see Fig. 6).

$$Y_1 = \text{Inci1} = \begin{cases} 1 \text{ si } TA > TC + 2\,^\circ\text{C} \\ 0 \text{ si } TA \leq TC - 2\,^\circ\text{C} \end{cases} \text{ and}$$

$$Y_2 = \text{Inci2} = \begin{cases} 1 \text{ si } TA < TC - 2\,^\circ\text{C} \\ 0 \text{ si } TA \geq TC + 2^\circ\text{C} \end{cases}$$

Table1 shows all the variables that will be used in the study. As can be seen, there are p = 30 covariates ($X_1, \ldots, X_{30}$) and two possible response variables (Y_1 *and* Y_2).

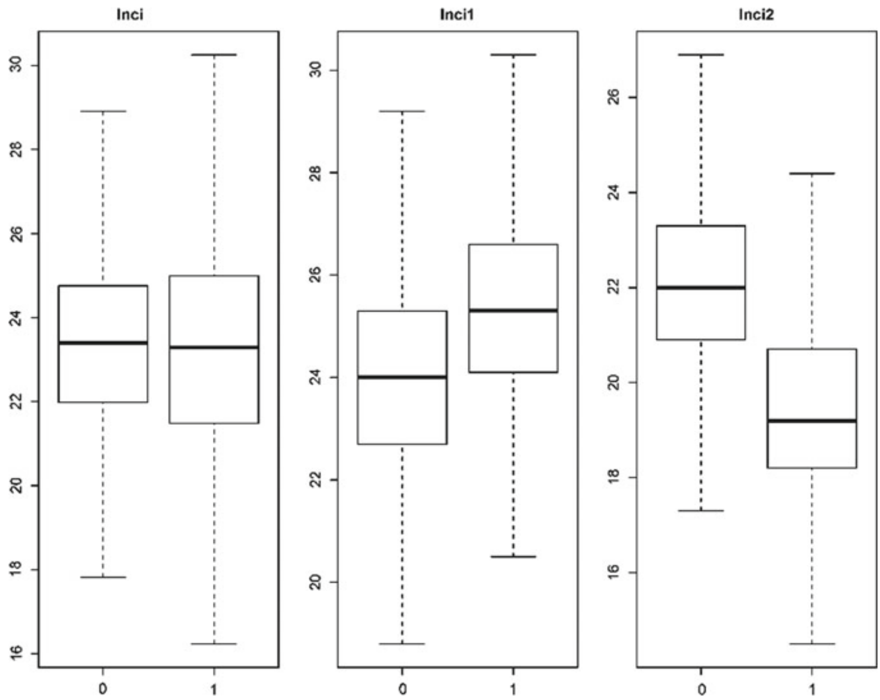


Fig. 6 Box diagram for the incident variables, Inci1 (in cooling regime) and Inci2 (in heating regime) taking the values 0/1 for the absence/presence of incidents, respectively

Table 1 Description of the variables

Notation	Variable name	Description
Y_1	inci1	Presence/absence of daily incidents in the cooling regime
Y_2	inci2	Presence/absence of daily incidents in the heating regime
X_1	amb.max	Maximum daily ambient temperature (closing time)
X_2	amb.min	Daily minimum ambient temperature (closing time)
X_3	amb.mean	Average daily ambient temperature (closing time)
X_4	amb.max.on	Maximum daily ambient temperature (opening hours)
X_5	amb.min.on	Daily minimum ambient temperature (opening hours)
X_6	amb.mean.on	Average daily ambient temperature (opening hours)
X_7	ext.max	Maximum daily outdoor temperature (opening hours)
X_8	ext.min	Minimum daily outdoor temperature (opening hours)
X_9	ext.mean	Average daily outdoor temperature (opening hours)
$X10$	gen.mean	Average daily general consumption (opening hours)
$X11$	gen.med	Average daily general consumption (opening hours)
$X12$	cli.mean	Average daily climate consumption (opening hours)
$X13$	cli.med	Average daily climate consumption (opening hours)
$X14$	iaire.mean	Average daily air supply temperature (opening hours)
$X15$	iaire.med	Average daily air supply temperature (opening hours)
$X16$	raire.mean	Average daily return air temperature (opening hours)
$X17$	raire.med	Average daily return air temperature (opening hours)
$X18$	iagua.mean	Average daily water flow temperature (opening hours)
$X19$	iagua.med	Average daily water flow temperature (opening hours)
$X20$	ragua.mean	Average daily water return temperature (opening hours)
$X21$	ragua.med	Average daily water return temperature (opening hours)
$X22$	con.mean	Average daily temperature setpoint (opening hours)
$X23$	horamin	Number of opening hours per day
$X24$	inci.5d	Number of incidents in the last 5 days including the current date
$X25$	City	City
$X26$	Type	Type of store
$X27$	Shop	Shop
$X28$	Machien	Machine
$X29$	Month	Month
$X30$	Day	Day

3.2.2 Prediction of Climate Active Power

The main objective of this project has been the definition of an algorithm that allows predicting the risk of incidents or lack of comfort in the premises. To do this, we have built a mathematical model that allows establishing relationships between a response variable of interest, Y, and a set, p, of explanatory variables $(X_1, \ldots, X_p)$.

In many practical situations, the response variable is qualitative or categorical, as presented in this work, where Y_1 (Inci1) or Y_2 (Inci2) can only take two possible values 0/1, (absence/presence of incidents in the cooling regime or heating regime, respectively). We have studied different approaches to predict qualitative response variables, a problem known as classification. In other words, the intention is to classify that observation by assigning it to a category or class. There are many possible classification techniques, or classifiers, that could be used in context. Some of the most used are regression models, discriminant analysis, k-nearest neighbors, random forest, decision trees, etc. (Hastie et al., 2003). Based on our previous experience, we have opted for the use of regression models.

3.2.3 Statistical Models: Generalized Additive Models

Generalized additive models (Hastie & Tibshirani, 1986) are flexible models which allow us to introduce non-linear effects of covariates, facilitating the interpretability and visualization of the relationship between the dependent variable and the predictor variables $X_1, \ldots, X_p$. In this case, we used a logistic GAM model with binary response $Y(non\text{-}targets = 0, targets = 1)$ given by

$$P(Y = 1|X_1, \ldots, X_p) = \frac{\exp\left(\alpha + f_1(X_1) + _{\ldots} + f_p\left(X_p\right)\right)}{1 + \exp\exp\left(\alpha + f_1(X_1) + _{\ldots} + f_p\left(X_p\right)\right)} \tag{1}$$

where α is a constant and f_p is the unknown smooth partial function or effect curve associated with each covariate X_p, estimated from the data. Both the introduction of a constant α into the model and th e requirement of a zero mean $E\left(f_p\right) = 0$ for the partial functions guarantee the identification of the model. The GAM is an extension of the traditional generalized models (GLM). They are flexible models since a parametric form for the effects of continuous covariates is not assumed, but instead, it is assumed that these effects may be represented by unknown smooth functions. Additionally, they are easy to interpret as the influence of each covariate is described separately by the additive components. To estimate this model, we used a procedure based on regression splines introduced by Wood (2004). According to this approach, each one of the smooth functions described in (1) can be denoted as:

$$f(x) = \sum_{i=1}^{k} B_i^m(x)\beta_i$$

where $B_1^m(\cdot), \ldots, B_k^m(\cdot)$ are a set of K known basis functions ($m + 1$ is the order of the basis function, e.g.,$m = 2$ for a cubic spline) and $\beta_1^m(\cdot), \ldots, \beta_k^m(\cdot)$, the associated coefficients estimated from the data. In order to define these B-spline basis functions (1978), it is necessary to define $k + m + 1$ knots, $x_1 < x_2 < \cdots < x_{k+m+1}$. Thus, the basis functions are defined recursively as follows:

$$B_i^m(x) = \frac{x - x_i}{x_{i+m+1} - x_i} B_i^{m-1}(x) + \frac{x_{i+m+2} - x}{x_{i+m+2} - x_{i+1}} B_{i+1}^{m-1}(x) \; i = 1, \ldots k$$

and

$$B_i^{-1}(x) = \{1 x_i \le x \le x_{i+1} \; 0 \; otherwise\}$$

We used the *mgcv* R package implementation of GAM, where the smooth functions are represented by penalized regression splines (P-splines). These are reduced rank smoothers that combine a B-spline basis- usually defined on evenly spaced knots- and a penalty applied directly to the basis coefficients β_i (Wood, 2006). By using the game function, given a sample $\{X_i, Y_i\}_{i=1}^n$ (n being the sample size and $X_i = X_{i1}, \ldots, X_{ip}$ the vector of explicative covariates), we can obtain the estimated components of the model (1): $\widehat{\alpha}, \widehat{f}_1, \ldots, \widehat{f}_p$.

3.3 Prediction Assessment

We used the receiver operating characteristic (ROC) curves (Swets & Pickett, 1982) in order to estimate the probability of false positives and false negatives in the prediction of a given model. By calculating the area under the ROC curve (AUC), we can measure the effectiveness of the prediction. The ROC curve is a plot of the sensitivity (Sen)—proportion of instances correctly classified as positives out of the number of true positives—versus 1-specificity (Esp)—proportion of instances correctly classified as negatives out of the number of true negatives—at all possible cut-points. The area under this curve measures the discriminative ability of a model. Two characteristics of this statistic are (1) it does not require any assumptions about the shape of the underlying signal and noise distributions (Saveland & Neuenschwander, 1990) and (2) it is a threshold-independent measure of model discrimination, the values of which range from zero to one, 0.5 suggesting no discrimination, 0.7–0.8 acceptable discrimination, 0.8–0.9 excellent discrimination, and >0.9 outstanding discrimination (Hosmer & Lemeshow, 1989).

To obtain the corresponding AUCs for the model (1), we used a k-fold cross-validation procedure, in particular $k = 5$, as follows. The sample is randomly divided into 5 groups (folds). For each subsample$k(k = 1, \ldots 5)$, $k - 1$ folds are used as a training set to learn the model. From this model, we obtain the probabilities for the k fold that is left out (test set) and the corresponding AUC_k. Additionally, given a particular cut-point c, we estimate the sensitivity and specificity values, denoted by $Sen_k(c)$ and $Esp_k(c)$. The values obtained with each fold $AUC_1, \ldots, AUC_k$ are then summarized by their mean$AUC = \sum_{j=1}^{k} AUC_k/k$. For each k, we also obtain a cut-off point using the Youden Index (Youden, 1950). This cut-off value is the threshold that maximizes the distance to the identity (diagonal) line, and therefore, optimizes the discriminatory ability of the classifier when equal weights are assigned toSen, i.e., $c_k = \left(Sen_k(c) + Esp_k(c) - 1\right)$. Again, the optimal cut-off point will be obtained as the average$c = \sum_{j=1}^{k} c_k/k$. Cut-off values were also estimated for fixed specificity values (0.90, 0.95 and 0.99).

3.4 Model Selection

The variable selection problem arises in the context of multiple regression models of type (1). The objective of variable selection is to determine the best subset of $q(q \leq p)$ predictors to include in the model in order to obtain the best predictive ability. A higher number of variables does not necessarily lead to better models for several reasons. Firstly, a large number of variables can make the resulting models difficult to interpret, in addition to being prone to collinearity and overfitting problems. Secondly, because of the bias-variance trade-off, the inclusion of irrelevant predictors would increase the variance of the estimated coefficients resulting in a loss of the predictive ability of the model and higher variability of a model prediction for a given data point.

In this study, we proposed a two-stage variable selection algorithm to select the best model to use for prediction, based on previously described procedures (Asís-López et al., 2014; Roca-Pardiñas et al., 2009; Sestelo et al., 2016). First, the best combination of q variables is selected using a step-by-step procedure. At the next stage, the number of predictors to be included in the model is then determined, based on a bootstrap procedure.

Thereby, in the first step, given a number $q(q \leq p)$ of predictors, the objective is to find the best combination of q variables. Let $AUC_{j1,\ldots,jq}(j_1 < j_2 < \ldots < j_q)$ be the AUC—computed as explained above—obtained by using exclusively the q covariates and omitting the remaining variables. Based on this metric, the best q predictors $X_{j1}, \ldots, X_{lq}$ can be selected. The vector of indices $\left(l_1, \ldots, l_q\right)$ is obtained by maximizing: $\left(l_1, \ldots, l_q\right) = AUC_{j_1,\ldots,j_q}$.

In the second step, once we have selected the best combination of q predictors, we need to determine the optimal number q of variables to be included in the final model. In order to do so, we test the null hypothesis of a model containing q variables versus the alternative hypothesis of a model containing $q + 1$ predictors. We denote this as

$$H_0(q)\colon AUC(q) = AUC(q+1)\ vs.\ H_1(q) : AUC(q) \neq AUC(q+1)$$

As in practice, given q, we do not know the value of $AUC(q)$, it must be estimated from a sample $\{X_i, Y_i\}_{i=1}^n$. The estimated $\widehat{AUC}(q)$ is obtained from the sample, following the cross-validation procedure explained above. Thus, when the difference $D = AUC(q+1) - AUC(q)$ is null, we will accept the null hypothesis $H_0(q)$: $AUC(q) = AUC(q+1)$. Accordingly, the unknown value of D must be estimated by $\widehat{D} = \widehat{AUC}(q+1) - \widehat{AUC}(q)$. Since $\widehat{D}$ is only an estimate of the true D, we should account for the sampling uncertainty of the estimates. Thus, we derived a $(1-\alpha)$, with $= 0.05$, simulation confidence interval $CI = (a, b)$ around D. We then check whether the 0 lies within this interval, which would imply equality of precision, or not, meaning that the models show different precision.

In order to estimate the distribution of $\widehat{D}$, resampling methods, such as the bootstrap (Efron, 1979) can be applied in the context of nonparametric regression. Thus, the bootstrap procedure implemented in this study is described as follows:

(1) For $= 1, \ldots, B(B = 1000)$, by randomly sampling with a replacement the original n dataset, simulate samples $\left\{X_i^b, Y_i^{*b}\right\}_{i=1}^n$. Thus, the bootstrap estimates $\widehat{D}^b$ are computed.

(2) Then calculate the $(1-\alpha)100\%$ limits for the simulation interval of D: $(\widehat{D}^{\alpha/2}, \widehat{D}^{1-\alpha/2})$, where $\widehat{D}^p$ represents the p percentile of the bootstrapped estimates $\widehat{D}^1, \ldots, \widehat{D}^B$.

By applying this bootstrap procedure, we can establish the model with the optimal number of predictors. Thereby, if $H_0(q = 1)$ is not rejected, given that the 0 lies in the bootstrapped confidence interval around D, the model will include only one variable. If this null hypothesis is rejected, it would be necessary to test $H_0(q = 2)$, and so progressively until a certain $H_0(q)$ is not rejected.

Apart from this type of iterative procedures, other strategies commonly used in the problem of selection of variables are, for example, regression shrinkage methods—such as LASSO (Least Absolute Shrinkage and Selection Operator) or ridge regression (Hastie et al., 2003), the methods based on trees—such as random forest, bagging or boosting (Breiman, 2001) or different Bayesian approximations.

4 Results

This section shows the results obtained taking into account the variables described in Table 1. In order to be able to study the probability of incidence in a certain store and to be able to predict the said phenomenon one day in advance, it is necessary to previously select the variables. As mentioned above, in the case of using a complete iterative procedure, it is necessary to construct $A = 2^p$ possible models, which implies that even for a moderate number p, making all possible subsets of predictors is intractable from a computational cost point of view (in this research $p = 30$, where $2^{30} = 1.073.741.824$). To solve this, we have used other techniques explained below.

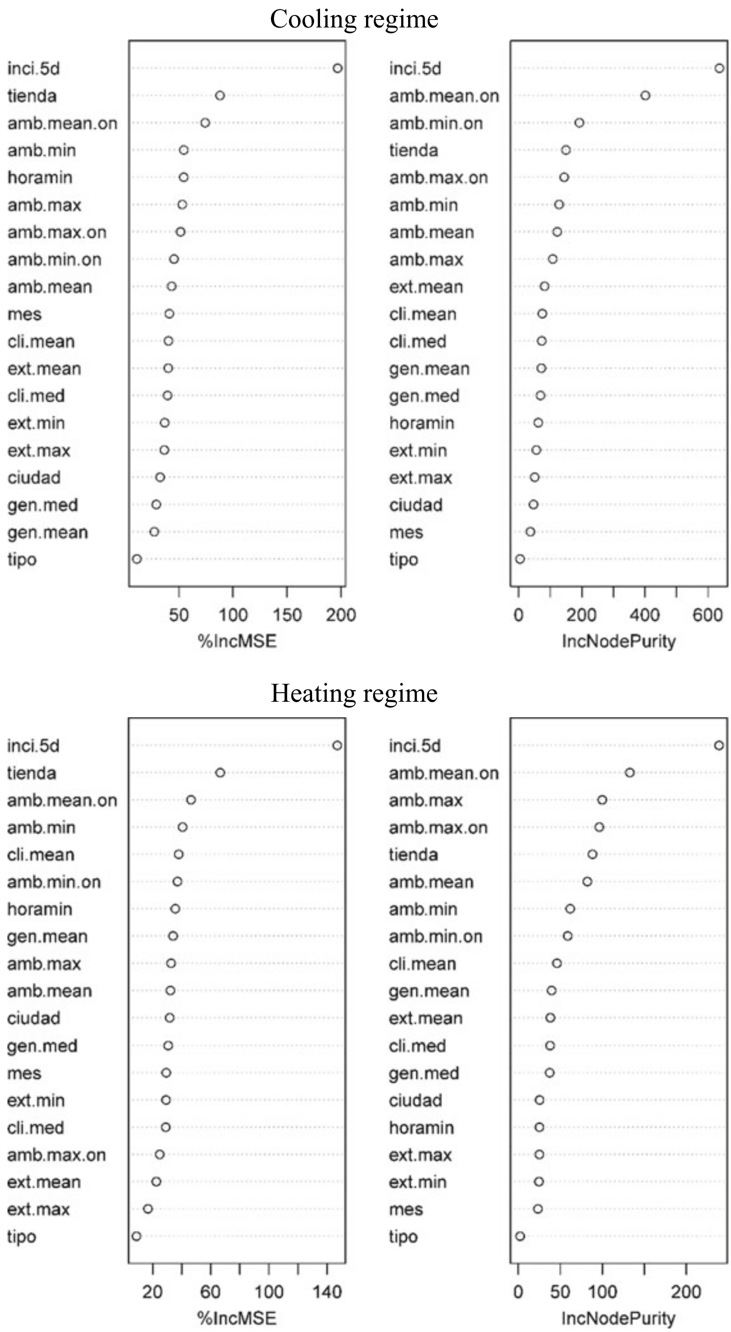


Fig. 7 Selection of variables for the cooling regime (top plots) and heating regime (bottom plots)

The results obtained are shown in Fig. 7, for example, with the random forest that allows determining a small subset of important variables. To carry out this selection, the procedure does not need to perform all possible combinations. As a result, it returns a plot with all the variables ordered according to their importance following two criteria. We have focused on the mean square error since it is the most used criterion in regression.

Based on the results, the subset of variables to take into account is maximum or minimum room temperature (during closing hours), average room temperature (during opening hours), average climate consumption, number of incidents in the last five days, number of hours the store was open, type of premises, store. In this way, we have reduced the set of variables $p = 30$ to only $p = 37$. Then, taking into account only this small subset of variables, all possible models have been constructed, $2^7 = 128$, and their corresponding AUCs have been calculated, both in the cooling and heating regimes (Fig. 8). In Fig. 8, and for each subset size, the models whose AUC is maximum are represented in red. The variables included in these models are shown in Table 2. In the cooling regime, the models with 2, 3, and 4 variables present similar AUC values, while in the heating regime the best AUCs correspond to models with 3, 4, 5, and 6 variables.

Taking into account the results shown in Table 2, we have selected the most parsimonious model, that is, the simplest model that allows us to obtain the best results. The models finally proposed are:

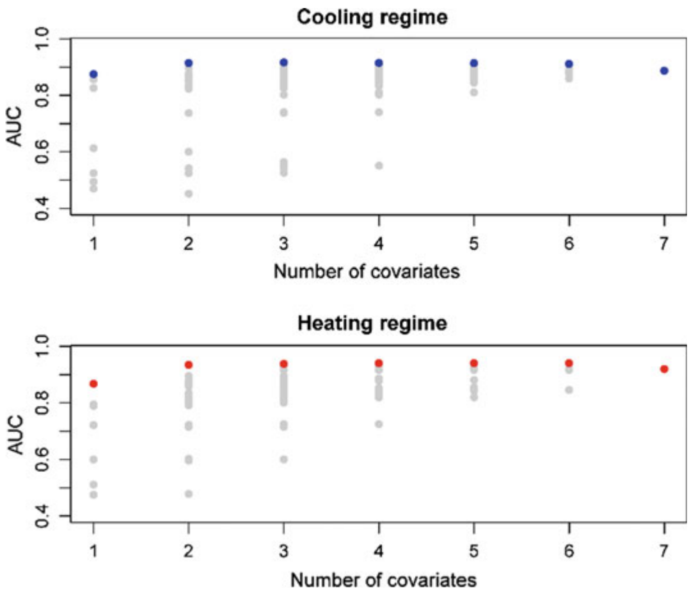


Fig. 8 All possible models using 7 covariates (in cooling and heating regime). For each covariate subset size (1–7), the AUC is shown in gray; the maximum value is represented in red

Table 2 Variables selected for the cooling and heating regime models

Cooling regime								
X1	X2	X3	X4	X5	X6	X7	nvar	AUC
amb.max							1	0.88
amb.max	inci.5df						2	0.92
amb.max	cli.mean	inci.5df					3	0.92
tipo	amb.max	cli.mean	inci.5df				4	0.92
tipo	amb.max	cli.mean	horamin	inci.5df			5	0.91
tipo	amb.max	amb.mean.on	cli.mean	horamin	inci.5df		6	0.91
tienda	tipo	amb.max	amb.mean.on	cli.mean	horamin	inci.5df	7	0.89
Heating regime								
X1	X2	X3	X4	X5	X6	X7	nvar	AUC
inci.5df							1	0.87
amb.mean.on	inci.5df						2	0.93
amb.mean.on	cli.mean	inci.5df					3	0.94
amb.min	amb.mean.on	cli.mean	inci.5df				4	0.94
tipo	amb.min	amb.mean.on	cli.mean	inci.5df			5	0.94
tipo	amb.min	amb.mean.on	cli.mean	horamin	inci.5df		6	0.94
tienda	tipo	amb.min	amb.mean.on	cli.mean	horamin	inci.5df	7	0.92

Table 3 Possible cut-off points for the cooling and heating model

Cooling							Heating						
Cut	TP	FP	FN	TN	Sen	Esp	Cut	TP	FP	FN	TN	Sen	Esp
0.0	283	3244	0	0	100	0	0.0	534	2993	0	0	100	0
0.1	229	454	54	2790	81	86	0.1	439	236	95	2757	82	92
0.2	221	275	62	2969	78	92	0.2	397	160	137	2833	74	95
0.3	204	127	79	3117	72	96	0.3	347	108	187	2885	65	96
0.4	188	68	95	3176	66	98	0.4	313	77	221	2916	59	97
0.5	174	41	109	3203	61	99	0.5	251	51	283	2942	47	98
0.6	124	28	159	3216	44	99	0.6	191	28	343	2965	36	99
0.7	106	8	177	3236	37	100	0.7	131	14	403	2979	25	100
0..8	37	2	246	3242	13	100	0.8	63	1	471	2992	12	100
0.9	0	1	283	3243	0	100	0.9	23	1	511	2992	4	100
1,0	0	0	283	3244	0	100	1.0	0	0	511	2993	0	100

Table 4 Confusion matrices for the cooling and heating models. The columns indicate the number of predictions for each class, while each row represents the instances in the actual class

Cooling			Heating		
	Predicted			Predicted	
True	0	1	True	0	1
0	2790	454	0	2757	236
1	54	229	1	1 95	439

- Cooling regime ("summer"): as response variable we use Y1 (Inci1), and as covariates, we include X1 (amb.max), X12 (cli.mean), and X24 (inci.5d).
- Model for the heating regime ("winter"): as response variable we use Y2 (inci2), and as covariates, we include X6 (amb.mean.on), X12 (cli.mean), and X24 (inci.5d).

The smoothing effect of the maximum ambient temperature for the cooling model is depicted in Fig. 9. It can be observed that as the temperature increases, the effect of this variable on the response also increases ("the probability of incidence" increases). This occurs until a certain temperature is reached, after which this effect stabilizes. The effect of the mean ambient temperature (during opening hours) on the "probability of occurrence" is opposite to the effect of the maximum temperature in the previous model. As the temperature increases, the risk of incidence decreases (see Fig. 9).

In our study, we have used regressors as classifiers, so prior to the construction of the matrix, it will be necessary to decide the probability or cut-off point from which we understand that there is incidence. As mentioned above, it must be taken into account that, depending on the cut-off point used, the false positive and false negative results

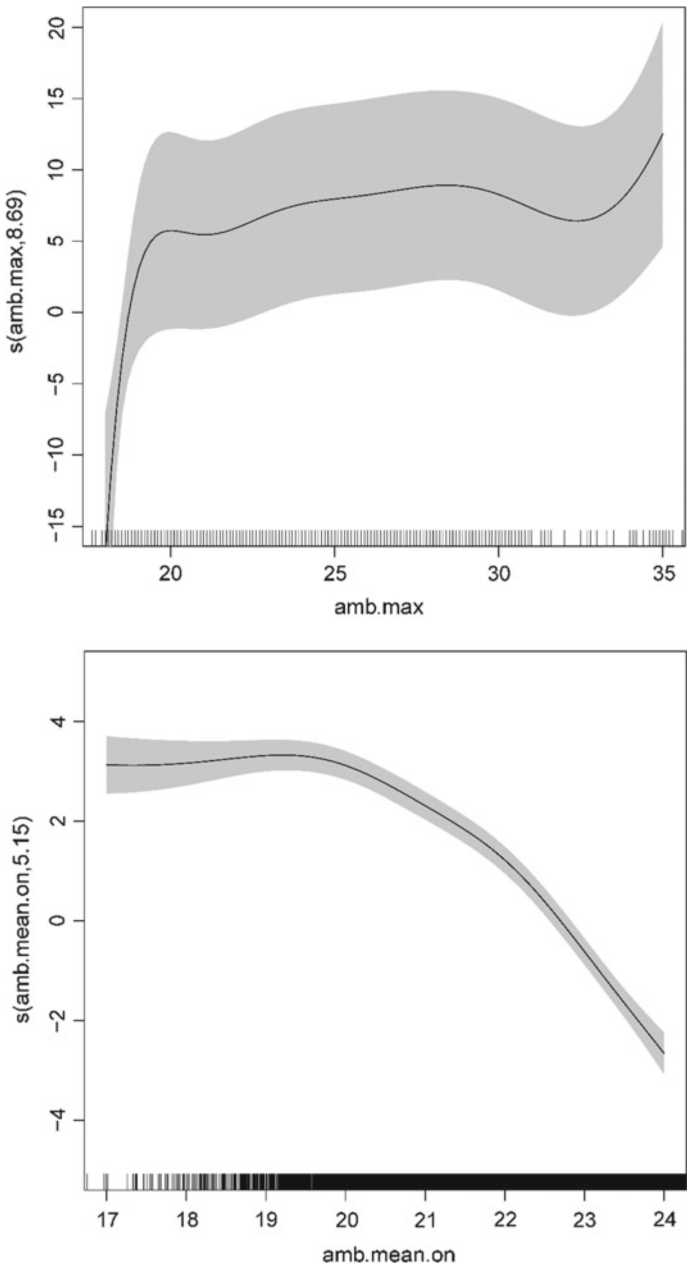


Fig. 9 Partial smoothed effect of maximum room temperature for the cooling regime (top plot) and mean room temperature for the heating regime (bottom plot)

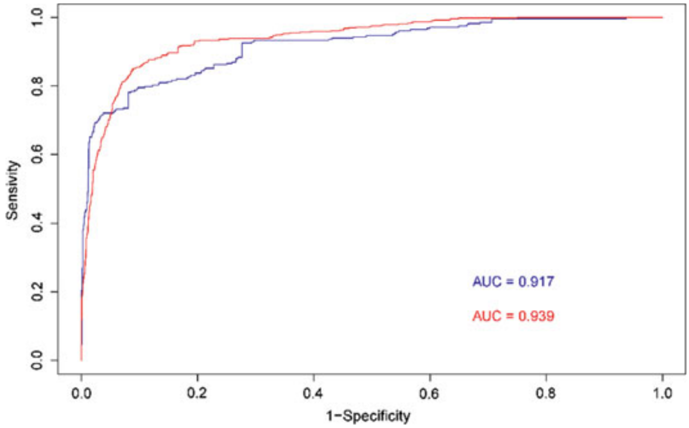


Fig. 10 ROC curves for the cooling model (blue line) and for the heating model (red line)

may vary. Next, the sensitivity (that is, true positive rate) and specificity (that is, true negative rate) values are shown for different cuts, both in the cooling model and in the heating model (see Table 3).

Next, the confusion matrix is shown for both scenarios studied (see Table 4), using a cut-off point of, for example, 0.5. These tables with two rows and two columns refer to the number of false positives, false negatives, true positives, and true negatives, and allow us to analyze the operation of the model in more detail.

Figure 10 shows the ROC curves obtained for both models. It seems that the warm-up model allows for better results. In any case, the AUC values are very close to unity, making them very adequate.

5 Conclusions

EcoMT aimed to detect incidents in advance through predictive maintenance of more than 3,000 installations. This is possible by applying ML techniques and the use of HPC.

For a client who uses the services of EcoMT, it is important to reduce incidents in a sustainable manner over time. To achieve this, it is essential to have an appropriate maintenance strategy, thus reaching an accurate diagnosis. After the application of the combined methods of ML and statistical techniques and their corresponding validation, it can be concluded that:

- The use of statistical and ML models is more effective than the deterministic system used by EcoMT before the experiment.
- Risk classification is useful to anticipate the appearance of incidences since high-risk situations are preceded by low-risk situations.

- A hierarchical algorithm helps classify incidences by their origin.
- These sorts of models and their application imply a steep learning curve.
- HPC makes it possible to test a large number of models in order to find the best. Once the best models are selected, training lasts a reasonable amount of time.
- Equipment failure situations that might compromise business activities can be prevented.
- An energy saving of 7% per installation is estimated, with a reduction of 30% in preventive maintenance visits and a reduction of 20% in corrective maintenance.
- More than 20% of the incidents registered could be corrected remotely without the need for a technician.
- Predictive maintenance will prolong the life of customers' HVAC machines and reduce the number that needs to be replaced.

The use of ML techniques and HPC opens new opportunities and presents improvements when managing the premises. By continuing to research these developments, EcoMT will add new lines of work, so that in the future, greater levels of efficiency and optimization in energy management will be obtained.

Acknowledgements This chapter describes a success story developed as an experiment under the Fortissimo 2 project which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 680481. The organizations involved were EcoMT, ITMATI, CESGA, and GOMPUTE.

Rubén Gayoso Taboada and Javier Roca Pardiñas acknowledge the support of ITMATI and EcoMT.

References

de Asís-López, F., Ordóñez, C., Roca-Pardiñas, J., & García-Cortés, S. (2014). Point cloud comparison under uncertainty. Application to beam bridge measurement with terrestrial laser scanning. *Measurement, 51*, 259–264.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1–26.

European Commission (2021a). Available at: https://energy.ec.europa.eu/topics/energy-efficiency/heating-and-cooling_en

European Commission (2021b). Available at: https://energy.ec.europa.eu/topics/energy-efficiency_en

Guide, C. D. B. A. P. (1978). A Practical Guide to Splines. Applied Mathematical Sciences, 27.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2003). *The elements of statistical learning*. Springer.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*, 297–310.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. John Wiley & Sons Inc.

Roca-Pardiñas, J., Cadarso-Suárez, C., Tahoces, P.G., & Lado, M.J. (2009). Selecting variables in non-parametric regression models for binary response. An application to the computerized detection of breast cancer. *Statistics in Medicine, 28*, 240–259.

Sestelo, M., Villanueva, N. M., Meira-Machado, L., & Roca-Pardiñas. J. (2016) FWDselect: An R package for variable selection in regression models. *The R Journal, 8/1*

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. Academic Press.

Saveland, J. M., & Neuenschwander, L. F. (1990). A signal detection framework to evaluate models of tree mortality following fire damage. *Forest Science, 36*, 66–76.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of American Statistical Association, 99*, 673–686.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Chapman and Hall/CRC

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*, 32–35.

Design and Development of a Discovery Service for Drivers Within the Connected Car Context Using Predictive Machine Learning Methods

Javier Goikoetxea and Alex Rayón

Abstract Currently, cars are equipped with a large number of electronic sensors that fulfil several functions. These vary from receiving and issuing a signal, to allow the automation through the permanent exchange of data and information. However, in this last field the decentralization of suppliers and manufacturers, and the lack of fully connected car solutions, have limited the creation of new solutions for both the drivers at a microeconomic level and for the general safety at a macroeconomic level. The way in which companies have developed their services to tackle these challenges have been through business rules. Historically, companies mixed up the geolocation of the vehicle with the proposals of the businesses with their own interests. This was a product-oriented approach, rather than a driver-oriented approach we propose. Additionally, we propose the usage of machine learning techniques that could scientifically show which activations are better to improve the value proposals for the drivers. Considering this context, we present a discovery platform for the drivers that could permit the recommendation of a service or a product when needed with the final focus of saving money. We also identify which variables are the most important ones in the maintenance and usage of the car. Considering a wide variety of variables, we show which ones explain better the behaviour of the drivers and show them ways to save money accordingly.

1 Introduction

In the last years, the automotive domain is changing. Among the different challenges the motor industry is facing is that vehicles are evolving from efficient engines into software machines. We are witnessing the beginning of a new era in the automotive industry where the concept of technology connected into a car arrives to completely

J. Goikoetxea · A. Rayón (✉)
University of Deusto, Bilbao, Spain
e-mail: alex.rayon@deusto.es

J. Goikoetxea
e-mail: javier.goikoetxea@opendeusto.es

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8_9

transform the vehicle. According to a study by PriceWaterhouseCoopers (2018), 40% of the mileage will be made by autonomous vehicles in the European Union by 2030. In that same year, the pool of vehicles is also expected to fall from 280 to 200 million. This represents a 28.5% reduction in the passenger car pool. This means a change of paradigm. Since there will be less demand for vehicles, they will have to last longer. It is also heading towards a world where the connected car is on the way to becoming a key car feature. An opportunity is therefore arising for the aftermarket world as fewer existing vehicles will have more activity and greater longevity is expected in those cars. Furthermore, according to the same report, vehicle manufacturers are gradually becoming software development companies to ensure that the needs of the autonomous car are met. While all this arrives, the world of the connected car is carving a niche in the evolution of this more autonomous society in terms of mobility.

This chapter aims to focus on this reality that the automotive industry is already experiencing today and that is beginning to change the management and treatment of the expenses that a person makes around the car. Twenty years ago, a vehicle had to be comfortable, safe, fast and fit with the fashion of the moment. Nowadays a car is something else. The vehicle has become a computer on wheels. This circumstance can be seen every time car manufacturers make a new commercial launch. In 1980, Spain had 10 million vehicles and just over 2,000 km of highways. At that time, the vehicles did not have any driving assistance system. The vehicles had to be efficient and beautiful. Today there are more than 30 million vehicles in Spain and more than 16,550 km of highways. The challenge lies not so much in waiting for manufacturers to develop new car models equipped with sufficient technology, but rather in embracing the opportunity to transform a conventional vehicle into a connected car through the installation of an electronic device into it. In Spain, there are over 30 million vehicles registered, out of which 25 million are simple cars. Of this figure, only slightly less than 5 million cars are less than 4 years old. So, we can assume that there is a potential market of almost 20 million cars aged over 4 years. The introduction of technology in this segment will help manage the car and its expenses more efficiently. As an average, it is calculated that in the EU a passenger car costs around €2,000 per annum (Grupo Next, 2019). Any variation below this would suggest significant savings for the drivers.

Within this context, we present NEXT Group (NEXT in advance) as a Spanish company, located at Madrid (www.gruponext.es). This company is focused on mobility data treatment and cost-efficiency driving models. NEXT has created a real-time communication platform with georeferenced and enriched data, to bring proposals and tailor-made solutions, depending on the end-user ontology. This company integrates with external systems (e.g. managers campaigns) and/or partners for data monetization by customized commercial campaigns. The most relevant use case of this platform is the Connected Car solution. The connected car solution generates and collects information from the car, using an *On Board Diagnostic* (OBD) connection. The focus of getting that data is to process and transform it into worth, based on big data analysis. All in all, the mission of the company is to generate and analyse relevant and monetizable information on the mobility ecosystem in order to provide ad hoc services to the end-user and B2B clients with both digital and physical experiences. The connected car platform solves the vehicle needs, when needed and only what is needed. It is relevant to note that nowadays that vehicles generate data that can be used to discover when a vehicle will need different services. Unfortunately, there is not a standard designed system to manage and collect car data. Aftermarket car services are numerous and can encompass actions such as refuelling, performing maintenance, parking the car or simply fixing a breakdown that has just occurred.

The way of turning a simple car into a smart car is via connecting a device into the car using three main components (see Figs. 1 and 2):

A. The Platform is a high-performance computer system, equipped with redundancy characteristics in all its critical elements, to allow the provision of the service without temporary interruptions. The platform also has interfaces with the information systems of the partners that provide real-time services to drivers.
B. The on board OBD device, connected to the car port of each of the vehicles, reads data from the various sensors that the car has, to monitor the status of its various components, to send them to the platform for processing and exposure (ordered and managed). Generically, we can call these data the telemetry of the

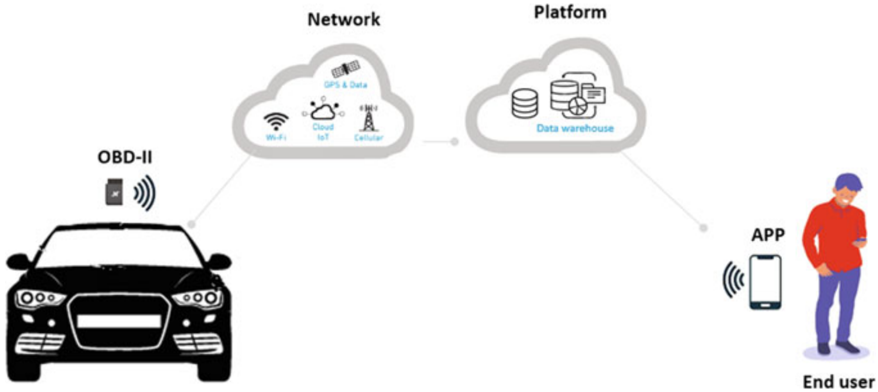


Fig. 1 High-level diagram of how the overall system

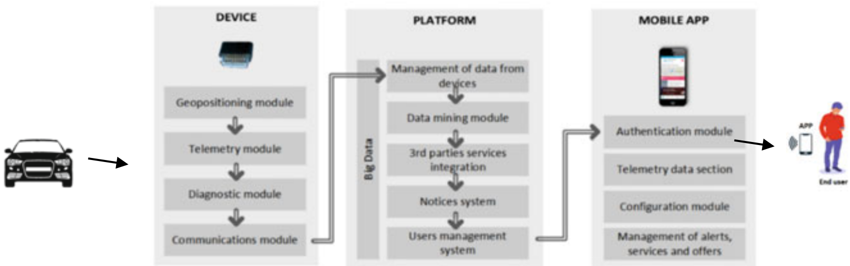


Fig. 2 Technical IT artefact solution high-level diagram

vehicle (Amouzegar & Patel, 2013). In addition, the device also incorporates a component to determine the geographical position in which the vehicle is using GPS technology. The most common technology of this type of device supports is formed by a GPS module, a Global System for Mobile communications (GSM) communications system, an accelerometer, a gyroscope, a small microprocessor, the diagnostic process and a backup battery (Duncan et al., 2015). The data will be collected from the OBD device that is part of our IT artefact, which will allow us to know what is happening in the car and its mobility in real-time and approach the user to propose a maintenance solution, when it is needed. DAS is a self-installing device which means that we will not have to use professional installers to board a device in a vehicle. This will simplify the testing. All this data is sent to a central computer to be analysed and processed normally thanks to the SIM allocated inside the artefact. The device contains a SIM card which, conveniently activated in the operator's 3G/4G network, facilitates the periodic sending of car info. The device is complemented by the creation of a communications protocol. This communications protocol will allow us to send to a server the device info.

C. A mobile application allows the driver to manage the system/service himself, from its activation to its deactivation, including its configuration, the subscription of new value-added services and some data visualization. It also offers information about the saved money in real-time, for car management purposes. This mobile application must be available for both Android terminals (versions equal to or greater than 4.4.x—KitKat) and iOS (versions equal to or greater than 9.0). The granularity of data collection as well as the frequency of sending them to the server can be handled manually by the operator of the artefact. According to Fogg (2009), the App, as end-user endpoint system, is a critical factor for the success of the system.

These three components consider technical aspects of efficient design. Like any element of automotive use, they must go through rigorous evaluation and testing methods to meet the required specifications (Resetar, 2016).

Currently, cars are equipped with a large number of electronic sensors that fulfil several functions from receiving and issuing a signal to automation through the permanent exchange of data and information. But the company platform we present is a full connected car solution. Coppola and Morisio (2016) define the connected car as '*a vehicle capable of accessing to the Internet, of communicating with smart devices as well as other cars and road infrastructures, and of collecting real-time data from multiple sources*'. This allows, for instance, that the car can communicate with the surrounding infrastructure (gas stations, workshops, spare parts suppliers, tolls…). The connected car will therefore affect the road network, the way of interacting with it, road safety and also the relationship between the vehicle and people. The vehicle, therefore, ceases to be a system that only transports us, to become a tool that can help us to satisfy our needs. It will cover them in real-time with the most accurate products and services available at that time. This technology must be able to control some parameters of the car and interact with the vehicle and its driver. Embedding technology in the vehicle is the catalyst for developing on-demand service models. From this point, it is very easy to control and determine how much a vehicle drives, how it is driven, what inspections need or what does it pay for petrol simply by parking the vehicle next to the gas station pump hose.

2 Problem Statement

With the previous section in mind, the idea is how to design a better line of services discovering when a service/product is needed to suggest a real-time solution opportunity to the driver with the final focus of saving money. In this sense, the problem can be studied from two sides: (i) variables affecting car consumption and (ii) the role of technology in reducing costs.

2.1 Variables Affecting Car Consumption

It is to be said that there are very few people who have dared to explore this terrain. Since the automobile has the second-largest share on the consumer´s durable goods pie (the first one is the house), the disbursements around it are also very relevant (Ferber, 1967).

Fuel consumption is one of the most relevant costs in-car use. Parry (2005) indicates in his article that he has estimated a reduction in fuel consumption of 9.1% for having embedded technology into the vehicle. It works as an element of persuasion and reminder, transforming drivers into more informed people and therefore more sensitive to spending.

It can be talked about car expenses but certainly not about the decisions taken beyond those expenses. Kim et al. (2018) indicate in their article that outgoings may vary depending on the type of driving, the driving circumstances, the type of car, the speed with which it is driven and other factors related to the car use. And this is true. All these factors affect the consumption of the vehicle but our research line highlights the consumer changes behaviour more than the effective car consumption.

However, there are other studies (Shanhan, 2019) that show that 44.33% of buyers of electric or hybrid vehicles are also buyers of home solar panels and 12.67% indicate that they have not yet bought the solar panels but they are going to buy soon. If both percentages are aggregated, it can be deducted that 57% of the people who drive a sustainable vehicle (*in environmental terms*) are also eco-driven in other activities of their ordinary life. This is a relevant factor since technology can help to condition the attitude of people and people who adopt an attitude as a pattern of behaviour, evolve according to that pattern (Fogg, 2009).

2.2 The Role of Technology in Reducing Costs

There is some literature that has focused on measuring some parameters of the car, just from a car activity control point of view (Amouzegar & Patel, 2013). These authors are focusing on car maintenance control through radiofrequency systems installed on roads. The ability to check a car remotely through technology (e.g.: change of oil, change of tires or change of brake pads) is part of what we are looking for in the literature review. According to car manufacturers, managing a vehicle on time has always an impact on an indirect cost reduction, since a possible deterioration is expected if this vehicle does not attend the revision on time. In this way, Lin et al. (2009) mentioned in their research that because of the remote on-line diagnostic system connected to a car, the time of fleet management and repairs can decrease significantly.

There is a current opinion that is looking at the impact of selling mobility instead of cars. According to empirical analysis from Firnkorn and Müller (2011), private vehicles were reduced as a result of a consumer reaction. This theory confirms that a vehicle is watched as an expense generator and there is, according to the same study, a part of users who want to pay only for their mobility.

Other studies have underlined the link between the use of technology and energy (Oppong-Tawiaha et al., 2020) and between the management of energy efficiency through the use of technology and gamification (Sousa et al., 2019). Until now, no writing evidence has been found in the scope of car use and savings management.

Apart from this, different authors point out that it is necessary to deepen in the analysis of the type of maintenance service that can be offered to a driver in real-time thanks to the adoption of new technologies (He et al., 2014). This call for research is focused on the management of the automatic calls for car revision (Lin et al., 2009), the collection of vehicle data for new services (Reininger et al., 2015) and the new business models for the automotive world (Jittrapirom et al., 2017).

In a nutshell, this chapter comes to fill in the gap of the inexistence in modern mobility literature of identifying which variables are most important about car expenditures and share the results in an empiric analysis of those variables.

3 Description of Previous Solutions

The way in which companies have developed their services to tackle these challenges has been through business rules. They have developed a system whereby the back end of the platform is able to collect different variables of mobility. These variables affect vehicle data as well as mobility data. The former refers to parameters of the car's engine (reading of levels, adequacy of the engine development, variables inherent to the control of the vehicle's electronic sensors, etc.) and the latter refers to the collection of mobility data (location, speed, vehicle movements in general, etc.). Before incorporating the description and analysis of the variables, all the data mining referring to its management and assimilation with real-time solutions is carried out based on business rules. These business rules have been defined by a set of criteria that mixed the geolocation of the vehicle with the definition of the business or the response that it was intended to launch towards the user.

Just to figure out how these business rules work, let's describe a real use case. Considering a situation when a specific vehicle passed a certain geographic point, regardless of its nature and its affinity for the product in question, the (business) rule was launched indiscriminately. In this sense, if a vehicle passed a certain kilometre point on a road, a small piece of information was sent to the vehicle, via the app, related to the advantages of refuelling at a particular gas station that was only a few kilometres from it was geo-localized. Thus, the rest of the variables necessary to determine the chance to send that precise opportunity to that vehicle were not considered. The commercial stimulus was sent only to meet the variable of the business rule: the occupant is passing by a certain geographic point.

This model for determining variables only by the description of a simple rule assumed that stimuli were being sent to customers just when a part of the business model was fulfilled, which was the drive-through a finished point. Thus, the need for creating more sophisticated algorithms to take into account other equally relevant variables to enrich the model was identified. This enriched model through an algorithm went from considering only the geographic variable to taking into account all possible variables. Elements such as the ontological needs of the drivers, the weather, the traffic density, the previous times in which they have stopped, the duration of those stops, the variable of the identification of the assignment of the suitability of the vehicle were considered. Another very relevant factor that could be analysed is the interest of that customer for the brand based on the affinity criterion of the social networks and other characteristics that went from identifying a simple model to a very sophisticated model where many other factors were taken into account.

In addition, we wanted to determine the ontological level of each user to use the factor of interests in intentions in the reformulation of the business rules. In this way, we wanted to go from using fixed variables to new variables that would help us to identify user purchasing trends. For this, we needed to incorporate a new methodology to develop the clustering of users.

On top of this, the company would like to know users' ontology. In order to do so, the company would need to understand what the behaviour of a user in sinister terms was. The company tried not to use the typical variables (speed depending on the type of road, number of kilometres travelled, smoothness or aggressiveness at the wheel …) but wanted to develop a new model that would allow us to determine new correlations to understand the behaviour of a user and their possible accident rate. Apart from this, the second main focus of this new approach to the end-user is to understand their interest, needs and intentions. All of those concepts are below the concept of ontology. For this reason, the company would like to develop a new system for treating the data. Up today, ontologies have become common on the internet world, to rank and categorize products and services for sale reasons (Noy & McGuinness, 2001). What the company is looking for is to analyse the users in order to prioritize their needs suggesting products and services in real-time, if needed and when those products are needed.

4 Solution Proposal and Results

The main objective of this work is to generate a model to predict, given a user whose mobility data we have, their ontological profile and, within this, their insurance risk profile (cars). We define the risk profile (SINCO score) of a user not as a value from 1 to 6 (which would force us to use regression methods), but as a class (class 1, class 2, … up to class 6 to which each user belongs). The risk profile will be trained against what is called the SINCO score, which is nothing more than the historical accident analysis shared by car insurance companies in Spain. It is a database that contains the accident rate and that allows car insurance companies to master the price of each of their car policies. For this reason, we call a classifier the model which can predict the class (SINCO score) based solely on the mobility data of a user.

In order to create and configure the infrastructure needed to capture data, its pre-processing, generate the customer´s ontology and finally the creation of subsequent models, we would need the following elements (see Fig. 3):

(1) Data capture and pre-processing: From a T3 request dedicated solely to the execution and monitoring of Python scripts, it is asynchronously, and multi-thread/multi-task initiated. We pre-process NEXT RAW data together with those of the client. We then add an incident and event notification system through AWS, SNS and Cloudwatch. All of these are preprocessed with AWS lambdas.

(2) User ontology (Data processing): The resulting data (list of users together with the categories of the POIs) coming from the pre-processing of the lambdas are then submitted to the Geo-Profiling Cluster to obtain the ontology for each user.

(3) Models: using Keras to control and execute the training in supervised and non-supervised models in Tensorflow, a cluster of 10 machines with dedicated GPUs has been built and configured.

In mathematical terms, a classifier occurs when we have input data (we call them ‘X’ and in our case they are the vectors with mobility data or independent variables), some output variables (we call it ‘Y’ and in our case they are the classes from 1 to 6 of SINCO, the dependent variable) and we use an algorithm to learn the ‘mapping function’ between the input data X and the output variables Y. This function is defined as $Y = f(X)$ and we call them classifiers or supervised predictive models. It is a classifier (or supervised model) because the process of an algorithm learning from a set of historical data (the SINCO results for each user along with their mobility data in the past) resembles the human learning process: the algorithm performs predictions iteratively over the historical data and is corrected until optimal performance is achieved.

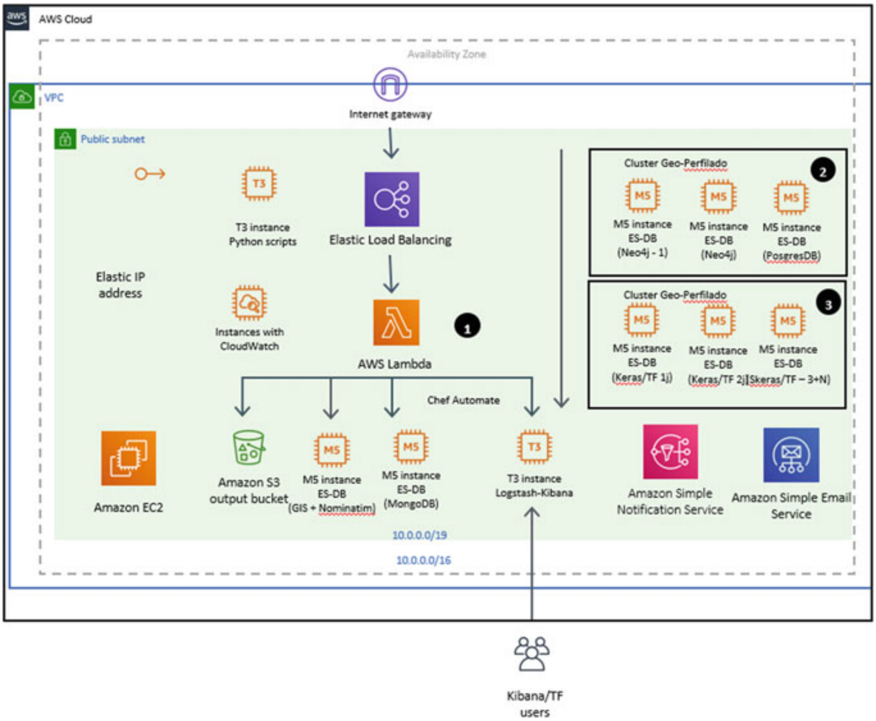


Fig. 3 The architecture of the proposed solution

The simplest algorithm to train a classifier consists of a linear function that separates each of the classes for all dimensions. The problem with simple linear classifiers is that, in real cases with many training data, many dimensions, and many classes, many possible solutions are presented with the same error. To solve this problem and at the same time achieve a function that better classifies classes that cannot be 'separated' in a linear way (as is the case with most of the real problems and in our case), we use the Support Vector Machines (SVMs) algorithm. SVM seeks the separation of the classes by maximizing the margin between the data closest to each other of each class (the 'support vectors') and applying on these a non-linear function that affects the entire class to which the' support vector belongs (see Fig. 4).

The mobility data of each user collected by the Next Group Company (www.gruponext.es) are data that we obtain from the GPS incorporated into the OBD-II device connected to the user's vehicle as explained in the previous points. These data are made up of the user's ID, longitude, latitude, vehicle speed, and the data collection time (timestamp) that the OBD-II device sends to the Grupo NEXT servers every 45 s. This data (see Fig. 5) is called raw data, since it is the primary data obtained from the capture device (the OBD-II device in the user's vehicle). In general, as in this case, it is data that individually, without further processing, does not contain

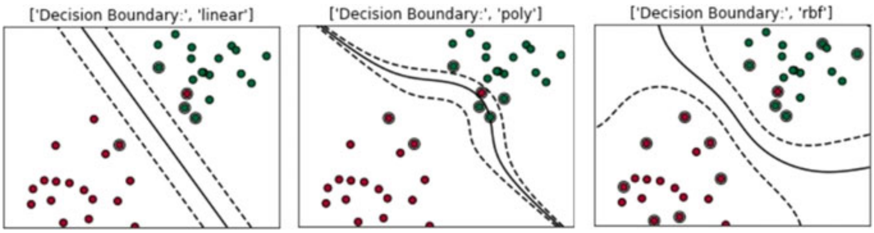
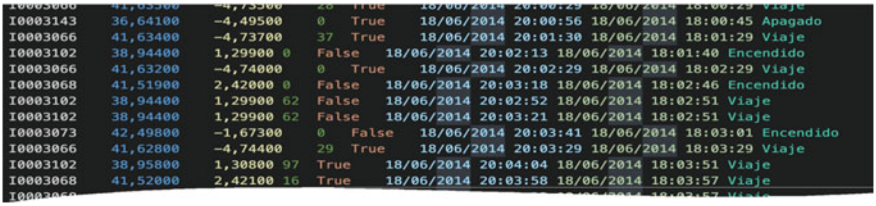


Fig. 4 SVM solution: **a** linear SVM; **b** Kernel SVM; **c** Radial SVM



I0003143	36,64100	-4,49500	0	True	18/06/2014 20:00:56	18/06/2014 18:00:45	Apagado
I0003066	41,63400	-4,73700	37	True	18/06/2014 20:01:30	18/06/2014 18:01:29	Viaje
I0003102	38,94400	1,29900	0	False	18/06/2014 20:02:13	18/06/2014 18:01:40	Encendido
I0003066	41,63200	-4,74000	0	True	18/06/2014 20:02:29	18/06/2014 18:02:29	Viaje
I0003068	41,51900	2,42000	0	False	18/06/2014 20:03:18	18/06/2014 18:02:46	Encendido
I0003102	38,94400	1,29900	62	False	18/06/2014 20:02:52	18/06/2014 18:02:51	Viaje
I0003102	38,94400	1,29900	62	False	18/06/2014 20:03:21	18/06/2014 18:02:51	Viaje
I0003073	42,49800	-1,67300	0	False	18/06/2014 20:03:41	18/06/2014 18:03:01	Encendido
I0003066	41,62800	-4,74400	29	True	18/06/2014 20:03:29	18/06/2014 18:03:29	Viaje
I0003102	38,95800	1,30800	97	True	18/06/2014 20:04:04	18/06/2014 18:03:51	Viaje
I0003068	41,52000	2,42100	16	True	18/06/2014 20:03:58	18/06/2014 18:03:57	Viaje

Fig. 5 Raw data available within the architecture of NEXT

much information. Therefore, its processing is necessary to obtain data with 'useful' information.

Raw data should not be used directly to train models (whether supervised or unsupervised classifiers). If we did so, what we would get is a classifier that would only be able to distinguish based on a given latitude and longitude or the closest point nearby to that corresponding to each class. For example, we would consider similar data the information coming from a user who works in the Jarama circuit (Madrid, Spain) or in a restaurant near it.

To obtain data that we can use to train a classifier, an inference process is necessary. In our case, starting from raw data representing geographic coordinates, we need to 'find out' which place, name it a school, company, store, shopping centre, highway, etc. (also called POI or Points of Interest) is at those coordinates. Once the POI of each coordinate is obtained and knowing the timestamp of each user (and the time that remains in each POI as well as the speed of the vehicle), we can infer the places visited by each user, how often and, based on this, draw subsequent inferences: where do you sleep, where do you work, if you take your children to school, the age of your children, etc. This process is called reverse geocoding. Due to the nature of the raw data (which is based on vehicle mobility data and not on the user), we encountered a problem when making the inference of which points of interest visited by the user: we knew where the vehicle was, but not the user.

To solve this problem and obtain the highest degree of confidence when assigning a POI to a user, we worked on various hypotheses based on the different POIs near the vehicle's parking place. For example, if a parked vehicle is 10 metres from 5 different stores, we assigned a low probability (0.01) to the potential visit of the user to each POI, since we cannot infer a visit to one store compared to another. In another example, we introduced the hypothesis that the recurring parking of a vehicle during working hours indicates the user's workplace, or the recurring parking of the vehicle during night hours the usual residence.

To obtain the data corresponding to each POI, it is necessary to use georeferenced databases, which provide us with the POI category (egg School, business, hospital, …) and within each category, the corresponding subcategories (egg School 'primary', 'electronics' trade, 'children's' hospital, …). This classification by categories and subcategories is called a taxonomy and each individual category is a taxon. In our case, we use a mix of OSM and Google Maps taxonomies:

Even though obtaining the relationship of POIs that a user has visited and the relationship of the user with each POI (e.g. 'Works in', 'sleeps in', 'has children of school age', etc.), it is then necessary to define the inferences that we are going to set for all users and how we are going to store and exploit them digitally. Or what is the same, what type of database we will use and how we will carry out the queries or to obtain the data that it is of interest for us from each user.

The set of inferences and data that we have predefined for each user is called the user ontology. This ontology or user profile (profile in its most complete and comprehensive state—not to be confused with a static profile that does not contain user relationships with each data), we store it digitally in graph databases. More specifically, it is a NEO4J graph database. To exploit this data (queries) we use the SPARQL language, which allows us to make queries using the relationships between each user with their data. For example, we can obtain all users who have school-age children and work less than 5 km from their school.

The process of selecting the data (dimensions or variables) and relationships in a user's ontology, we call it 'dimensionality reduction'. If N is the number of dimensions or variables that we have obtained for the user, in our case it will be all the different categories of POIs, in addition to the set of all possible relationships (we call it R) between them and the user (the relationships of the ontology). The real number of possible dimensions will be the Cartesian product of $N \times R$. In our case we have an N of 10,200 (number of taxa) and an R of 1,100 (total of possible relationships defined in the ontology), then N (the overall dimensions) will be 11,220,000. If M is the number of users with data and Y results (understood results as a SINCO class of the history and Y the 'true' observations of the function $Y = f(x)$ that we saw previously), it leads us to the maximum number of Users with useful observations to train a supervised model will be, in our case, 40,500 (the number of common users).

The problem we encounter is that (mathematical proof aside), when N is much greater than M ($N >> M$), any supervised classifier that is trained by any algorithm will result in unstable models with low performance (the so-called the 'curse of dimensionality'). So, we need M to be much greater than N ($M >> N$). Even if we eliminate the Rs (that is, we assume that in the ontology all relations are of the type '*the user is interested in a certain POI category*'), we would still have an N of 10,200 and M of 40,500. The only option is to reduce N to those dimensions that contain more discriminatory information between the different classes. Out of the dimensionality reduction techniques, the most developed one and with the best results for the type of data we have is the Principal Component Analysis (or PCA according to its acronym). This process is applied to all variables and with this we obtain an N″ with sets of 30–100 dimensions, which are the ones we proceeded to train the supervised models using SVMs.

The PCA technique consists of extracting the normal values out of all the observations and all the dimensions, in such a way that when sorting them from highest to lowest, we obtain for each dimension the ‘information provided by each dimension in the discrimination of each class’. Now we only need to select which of these will be the ones we will use to train our supervised classifier. There are two techniques. One is to select the dimensions manually by an expert when the domain gives rise to it (not in our case, since it is difficult to know which dimensions of an ontology are important for SINCO beforehand). The second is automating it by generating so many supervised models using sets of them (30, 40, 50… sets of variables) and having them compete with each other to obtain the optimal number of dimensions.

This process (together with training using SVMs) is the process with the highest computational cost, since the number of models that compete with each other will ideally be formed by the Cartesian product of the size of the set of selected variables (the set T of sets 30, 31, 32, ... up to N″/2) and the number of possible dimensions (10,200). In order to avoid a computationally unmanageable number of classifiers, we have decided to reduce the number of sets to 150 experiments (and it is still 3 days of computational time for each PCA process plus dimension selection)

Once we obtain a trained classifier, we must be able to test it with users who have not been used in the training process. To do this, the first thing to focus on is the creation of a subset with the N data, we call it N″’, which consists of 33% of the users with SINCO data. This is our test data set for each model generated. This causes N to be reduced (from 40,500 to 30,000) and N″ from 10,500.

The results of the classifier on the test data can then be clustered as follows:

- True Positives (TP): when the real class of the data point was 1 (True) and the predicted class is also 1 (True).
- True Negatives (TN): when the actual class of the data point was 0 (False) and the predicted class is also 0 (False).
- False Positives (FP): when the actual class of the data point was 0 (False) and the predicted one is 1 (True).
- False Negatives (FN): When the actual class of the data point was 1 (True) and the predicted value is 0 (False).

If we group this data in a matrix, we obtain for each classifier a matrix that we call a ‘confusion matrix’.

Starting from the confusion matrix, we can then obtain the following metrics:

- Precision (what proportion of positive identifications have really been correct?) = TP/(TP + FP).
- Recall (what proportion of real positives has been correctly identified?) = TP/(TP + FN).
- Accuracy (which of the predictions did our classifier correctly identify?) = (TP + FN)/(TP + TN + FP + FN).

Establishment of the processes for the evaluation of supervised and predictive models (aimed to replicating the SINCO risk score and developing the complete user ontology):

(1) Extraction of 'blind' customers for independent evaluation—1.000 blind customers have been drawn (with whom no training nor internal evaluation in any model have been carried out) out of the 40.500 common customers (from the client and as from NEXT). Additionally, two other groups were created: one comprising 30,600 customers to train the models and another one of 6,000 customers to internally evaluate the models

(2) As a specific request from the client, some other variables on the type of customer driving together with the variable of temporality have been included (…). These variables have been incorporated into the models to be trained again, forcing their crunching regardless of their weight in the discrimination of the data, in order to evaluate them against the rest of the variables in depth. Time dimensions have also been incorporated into current variables

- Total driving time—the sum of a customer's total driving minutes on any type of road. It is defined as the time with the vehicle at a speed >0 and normalized by the total time as a customer.
- Total high speed time—It is defined as the total time with the vehicle at a speed >100 km/h of a customer on any type of road. It is then normalized for total time at speed >0.
- ($\times$3) Total time on Highway, Secondary and Regional/Other roads—total time with the vehicle at a speed >0 km/h for each type of road. Normalized by the total driving time.
- Total time of violations—total time with the vehicle at a speed 20% higher than that allowed on any type of road.
- Total number of infractions—number of trips, out of the total made, with at least one infraction committed for speeding.

(3) Analysis of context, events and meta-data—No customer's context has been incorporated at any time nor the meta-data corresponding to unique events or situations. For example, special events such as concerts or demonstrations, which give us information about the customer's profile, have not been linked to geolocations.

Of all the variables used, we are left with those with the highest incidence on the data, ranked from the highest to the lowest number of events. We highlight with a red dot those that have a disproportionate weight over the rest and on which we will include a third level of taxonomy (e.g. (Amenity, Pub) (Amenity, Pub, 00:00 to 06:00 LV, NOT Work_Place)) (see Fig. 6).

	office	estate_agent		highway	residential		shop	lottery
	amenity	hospital		shop	tobacco		office	ngo
	leisure	golf_course		shop	hardware		shop	greengrocer
	shop	daityourself	x	tourism	caravan_site		office	telecommunication
	railway	platform	x	shop	garden_centre		building	train_station
	toursim	artwork		amenity	post_office		shop	gift
	historic	monument		amenity	police		building	school
	shop	sports		amenity	marketplace		shop	jewelry
	amenity	theatre		highway	proposed		building	roof
	military	checkpoint		amenity	library		shop	car_repair
x	amenity	pub		historic	memorial		amenity	training
	tourism	attraction		highway	construction		amenity	vehicle_inspection
	amenity	public_building		amenity	bicycle_parking		amenity	vending_machine
	amenity	parking		shop	clothes		amenity	waste_transfer_station
	craft	window_construction		natural	spring		building	church
	office	company		amenity	language_school		shop	convenience
	amenity	social_facility		building	commercial		highway	primary_link
	building	appartments	x	amenity	embassy		building	dormitory
x	leisure	marina		building	garage		shop	motorcycle
	leisure	playground		place	locality		tourism	information
	office	yes		railway	station	x	military	barracks
	amenity	bank		highway	footway	x	military	bunker
	amenity	fuel		amenity	university		natural	heath
	shop	mall		historic	ruins		natural	water
	shop	tyres		highway	track	x	amenity	kindergarten
x	amenity	brothel		building	terrace		amenity	charging_station
	leisure	stadium		amenity	townhall		shop	variety_store
	natural	beach		leisure	garden		shop	wine
	shop	furniture		building	industrial		office	foundation
	tourism	hotel		club	culture		amenity	exhibition_centre
	shop	bakery		office	government		shop	pet
	highway	service		aeroway	aerodrome		shop	music
	amenity	cafe		highway	services		shop	trade
	amenity	bar		highway	cycleway		office	educational_institution
	highway	trunk		leisure	pitch		railway	subway_entrance
	highway	primary		shop	kiosk	x	shop	bicycle
	shop	hairdresser		shop	optician		building	house
	building	yes		place	square		building	garages
	amenity	bus_station		amenity	clinic		office	it
	amenity	pharmacy		amenity	college		historic	castle
	highway	secondary		man_made	surveillance		building	hospital
	place	house		building	office		office	insurance
	tourism	museum		highway	living_street		shop	car_parts
	highway	pedestrian		shop	department_store		leisure	common
	amenity	restaurant		amenity	fast_food		tourism	guest_house
	leisure	sports_centre		amenity	motorcycle_parking		amenity	bicycle_rental
x	amenity	place_of_worship		amenity	community_centre		tourism	camp_site
x	highway	tertiary		highway	motorway		shop	yes
	shop	supermarket		shop	car		shop	electronics
	highway	unclassified		leisure	swimming_pool			
	amenity	school		building	residential			

Fig. 6 Variables that have a disproportionate weight over the rest

5 Results Evaluation and Conclusions

We have presented the variables that, out of over 1.500 variables analysed (without including the variables with time dimension), best represent the behaviour of the driver. Evaluation and influence of these variables in the models has also been presented, being the best the one that presents an accuracy of 92.6% and a recall of 74.1%. The 150 variables with the highest incidence on data have been ranked from upper to lower level of events. We have highlighted with a red dot those that have a disproportionate weight over the rest and on which we will include a 3rd level of taxonomy.

A total of 900 computing hours have been conducted to date (in 53 uninterrupted days) for data processing and generation of supervised models. We have incorporated the time dimension variable into modelling, but with some limitations. For example, when a customer is driving on a county road (day or night) or the hours and days of the week when a customer visits a Pub or a Restaurant (weekend vs. midweek). Perhaps more granular time variables could be incorporated.

Apart from that, from a business point of view, we could also include variables to analyse users from an economic perspective: higher margin contribution, greater capacity for cross-selling and up-selling and last but not least, less churn. That could provide NEXT a further step in improving its' business value proposition.

References

Amouzegar, F., & Patel, A. (2013). Vehicle maintenance notification system using RFID technology. *International Journal of Computer Theory and Engineering,* 312–316.

Coppola, R., & Morisio, M. (2016). Connected car: Technologies, issues, future trends. *ACM Computer Surveys, 49*(3), 1–36.

Duncan, M., Charness, N., Chapin, T., Horner, M., Stevens, L., & Richard, Y. A. (2015). *Enhanced mobility for aging populations using automated vehicles.* BDV30 977-11

Ferber, R. (1967). Determinants of investment behavior. In *NBER books, national bureau of economic research*, Inc, number ferb67–1.

Firnkorn, J., & Müller, M. (2011). What will be the environmental effects of new free-floating car-sharing systems? The case of car2go in Ulm. *Ecological Economics, 70*(8), 1519–1528.

Fogg, B. J. (2009). A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology (Persuasive '09).* Association for *Computing Machinery.* New York, NY, USA, Article 40, 1–7.

He, W., Yan, G., & Xu, L. D. (2014). Developing vehicular data cloud services in the IoT environment. *IEEE Transactions on Industrial Informatics, 10*(2), 1587–1595.

Jittrapirom, P., Caiati, V., Feneri, A. M., Ebrahimigharehbaghi, S., Alonso González, M. J., & Narayan, J. (2017). *Mobility as a service: A critical review of definitions, assessments of schemes, and key challenges. Smart Cities – Infrastructure and Information, 2*(2)

Kim, J., Hwangbo, H., & Kim, S. (2018). An empirical study on real-time data analytics for connected cars: Sensor-based applications for smart cars. *International Journal of Distributed Sensor Networks, 14*(1), 1550147718755290.

Lin, J., Chen, S., Shih, Y., & Chen, S. (2009). *For vehicles by integrating the technology of OBD, GPS, and 3G* (Vol. 56). Word Academic of Science.

Noy, N. & Mcguinness, D. (2001). Ontology development 101: A guide to creating your first ontology. *Knowledge Systems Laboratory, 32.*

Oppong-Tawiah, D., Bassellier, G., & Pinsonneault, A. (2020). Tracing the next-generation platform firm: A typology of digital platforms as new organizing forms. *ECIS 2020 Research-in-Progress Papers,* 9.

Parry, I. W. H. (2005). Is pay-as-you-drive insurance a better way to reduce gasoline than gasoline taxes? *American Economic Review, 95*(2), 288–293.

PricewaterhouseCoopers. (2018). *En 2030, Europa tendrá 80 millones de coches menos como consecuencia del transporte compartido y de la digitalización.* Available at: https://www.pwc.es/es/sala-prensa/notas-prensa/2018/2030-europa-ochenta-millones-menos-coches.html

Reininger, M., Miller, S., Zhuang, Y., & Cappos, J. (2015). A first look at vehicle data collection via smartphone sensors. In *2015 IEEE Sensors Applications Symposium (SAS)* (pp. 1–6).

Resetar, M. (2016). *Innovative approach to vehicle diagnostics.* Available at: https://docplayer.net/33754531-Innovative-approach-to-vehicle-diagnostics.html.

Shahan, Z. (2019). EV Ownership + Rooftop Solar Ownership—New Report & Charts. *CleanTechnica.* Available at: https://cleantechnica.com/2019/12/25/ev-ownership-rooftop-solar-ownership-new-report-charts/.

Sousa, S., et al. (2019). Gamification as a way to involve young adults in energy efficiency and sufficiency—A case study. *ECEEE Summer Study Proceedings.*

Index

D. Carou et al. (eds.), *Machine Learning and Artificial Intelligence with Industrial Applications*, Management and Industrial Engineering,
https://doi.org/10.1007/978-3-030-91006-8