INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Applied Data Science on Children Health & Well-being

João Rala Cordeiro

PhD in Information Science and Technology

Supervisor:

Professor Dr Octavian Postolache, Associated Professor

ISCTE-IUL, Instituto de Telecomunicações

December 2022

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology (ISTA)

Applied Data Science on Children Health & Well-being

João Rala Cordeiro

PhD in Information Science and Technology

Júri:

Prof. Francisco Cercas, Full Professor, Iscte-IUL (President)
Prof. Francisco Ferrero, Associate Professor, University of Oviedo
Prof. Rui Madeira, Associate Professor, Polytechnic Institute of Setúbal
Prof. Sancho Oliveira, Associate Professor, Iscte-IUL
Prof. Octavian Postolache, Associate Professor (with Habilitation), Iscte-IUL

December 2022

Dedication

The idea for this thesis arose from the desire to develop my career and use my experience to benefit the community, contributing in a more significant way to society. The possibility of becoming a father heightened this desire.

Today, the "little sprout" who inspired the decision is already born and, at 12 months old, is bringing joy to the entire family. It is for him, in particular, the dedication of this thesis.

Romeo, my sun, is the personification of all children worldwide, and I dedicate this work to them in the hope that I can make a positive contribution for their future.

Acknowledgements

When I made the decision to begin this PhD Thesis, I knew it wouldn't be an easy task. To be completely honest, I knew that I would start it, I was sure that the challenge would help me develop both personally and professionally, and that by the time it was over, I would have acquired new skills and knowledge. However, I was not sure whether it would be possible to complete the entire journey.

At this point, where I'm nearing the end of the journey, I fully realise that it was only possible because I had the support, collaboration and help of many people.

Firstly, I have to express my gratitude to Professor Dr Octavian Adrian Postolache who, from the beginning, was willing to listen to my ideas and thoughts, which at first may have seemed a little crazy. Prof. Dr. Octavian provided full support for this PhD thesis in a variety of ways, including through his patience, inspiration, ideas, creative problem-solving, and vast knowledge.

Professor Dr. João Carlos Ferreira was also a very important person in the development of this Thesis. In some moments, when the motivation was quite low, at least in two key moments, it was Prof. Dr. João Ferreira who helped me find the strength to continue this difficult journey, by supporting in finding solutions to the issues I needed to immediately solve. I'm deeply grateful for his help.

The INOV INESC group also deserves my sincere gratitude for providing me the chance to find some personal funding, which allowed me to continue this research. Nevertheless, this group provided me with much more than just financial support; it also gave me the opportunity to advance my data science knowledge and broaden my knowledge while working with a talented team. In particular, I'd like to thank Pedro Santos and António Leal for the opportunity, Filipa Soares Borrego, Ricardo Ribeiro, and the entire team for being such great colleagues and always being ready to lend a hand and support an idea.

A Data science thesis obviously requires data, it is a crucial component. Regarding this subject, there were some people who were truly essential to achieving the proposed goals. A very special thank you to Prof. Dr. Maria Teresa Neto, from NOVA Medical School and Direcção-Geral da Saúde (Portuguese government) for the support in the partnership with DGS and for all the research work we have accomplished together. Another fantastic and crucial collaboration was carried out with Geração XXI project. Professor Dr. Henrique Barros, Faculdade de Medicina do Porto and Instituto de Saúde Pública da Universidade do Porto, was extraordinary in believing in our ideas from the very beginning, and in his willingness to "open doors" for the opportunity to develop science together with Geração XXI team. To the Geração XXI, and specially to Prof. Dr. Liane Costa I must express my deepest gratitude. Liane Costa was absolutely tireless and always willing to help.

While developing this thesis I had the opportunity to join the Instituto de Telecomunicações research group. I appreciate how the group have included me in the team. It was really stimulating the opportunity to discuss several subjects, about research, innovation, among other topics. I'm particular grateful to António Raimundo for the opportunity to exchange ideas and for the support with some of the thesis initiatives. It was really important to have his support.

The organisation Instituto de Telecomunicações itself was also crucial. It has provided all of the technological resources and technical support that made it possible to accomplish the thesis' objectives.

Finally, but certainly not least, I want to thank my family and friends. My sun which, although being very young, was a great contributor with good nights of sleep and an always-ready smile. To Sara, my partner, who has always supported me through good and bad times, to my parents, especially my mother, who has always - and I must emphasise always - had supported me in developing my academic potential. To my friends and closest persons, thank you for listening to my "complaints" and offering advice, and also for being there for me to laugh and enjoy good times.

This research work was partially funded by Fundação para a Ciência e Tecnologia (Ph.D. Research Scholarship reference 2020.07443.BD). I would like to thank Fundação para a Ciência e Tecnologia for this fantastic opportunity.

Resumo

Todos os anos são gerados, em Portugal e no mundo em geral, enormes quantidades de dados relativos à saúde e ao dia-a-dia de um individuo. Desde os dados mais básicos como são exemplo o peso ou a altura, a informações decorrentes de situações de doenças, toma de fármacos ou sintomas, ou ainda resultados de diversos meios de diagnóstico complementares, como são exemplo as análises clínicas ou a ressonância magnética. Esses dados são tipicamente utilizados uma ou duas vezes, numa observação clínica e posteriormente guardados para não mais serem utilizados.

Esta tese teve como objetivo investigar se é possível extrair valor e conhecimento a partir de *datasets* pré-existentes, obtidos e compilados em contextos de saúde, mas tendo como perspetiva outros projetos que não de ciência de dados. O derradeiro objetivo é o contributo para a saúde e bem-estar da nossa sociedade, começando desde logo pelas crianças. A tese concluiu que é de facto possível, recorrendo para isso a várias técnicas de data science e machine leaning entre as quais redes neuronais convolucionais, algoritmos de ensemble, inteligência artificial "explanibility", pesquisa de arquiteturas de redes neurais entre outras técnicas. Por vezes existiu a necessidade de criar novos recursos (inexistentes na comunidade científica) para alcançar os objetivos propostos. Esses recursos foram posteriormente partilhados com a comunidade.

Não obstante os recentes avanços na investigação nas áreas de machine learning e ciência de dados, e da abundância e (ainda em) expansão de recursos, a maioria dos projetos – se não todos – são concebidos a partir do zero ou com base em "datasets científicos". Os dados existentes na área da saúde representam, na nossa opinião, uma vastidão de recursos subutilizados.

A tese explorou e demonstrou a capacidade de derivar conhecimento e valor científico a partir de dados existentes, apresentado inclusive três casos de uso reais. Em todos casos foi possível endereçar questões científicas pertinentes e relevantes, como sejam as alterações cardiovasculares causadas pela obesidade infantil ou os fatores que influenciam as infeções neonatais hospitalares analisados nas dimensões de relevância e impacto.

Durante esta investigação, foram criados e partilhados com a comunidade vários recursos. Estes recursos incluem a proposta de uma nova metodologia focada em ciência de dados na área da saúde, várias técnicas para explorar conjuntos de dados existentes, novas funcionalidades e expansões em bibliotecas de ciência de dados (ex: auto-skearn ou AutoKeras), ou métodos para otimizar arquiteturas de machine learning.

A equipa de investigação tem a expectativa de que esta tese possa impulsionar novas iniciativas de ciência de dados no campo da saúde, inclusive em áreas não relacionadas com crianças.

Keywords: Ciência de dados na saúde; eXplainable Artificial Intelligence; NAS e optimização de arquitecturas; Técnicas de ciência de dados para saúde; Contribuições para melhorar a saúde das crianças

Abstract

Every year in Portugal and generally all over the world, enormous amounts of data are produced about a person's health and daily activities. Information resulting from disease situations, drug use or symptoms, the outcomes of various complementary diagnostic methods, like clinical analyses or magnetic resonance imaging, or simply data like weight or height. The typical usage pattern for such data is for one or two clinical observations, after which it is stored with no further use.

This thesis sought to ascertain whether knowledge and value could be extracted from data that had previously been gathered in health contexts but not with data science projects in mind. The ultimate goal was to contribute to our society's health and well-being, starting from young ages. It was determined by the thesis that it is indeed feasible, using several data science and machine leaning techniques, such as convolutional neural networks, ensemble algorithms, explainable artificial intelligence, neural network architecture search and other advanced techniques. To accomplish the goals of the thesis, it was necessary to develop new resources (non-existent in the scientific community), which were then made available to the scientific community .

Despite the recent advances in machine learning and data science research and the abundance of several and expanding resources, the majority of projects—if not all—are design from scratch or based on “scientific” datasets. The existing health data is, in our opinion, a vastly underutilised resource.

The thesis has explored and demonstrated the ability of deriving scientific knowledge and insights from the available data, inclusive presenting three real-world use case. In all cases, pertinent and important scientific issues were addressed, such as the cardiovascular changes caused by childhood obesity, or the factors influencing neonatal hospital infections analysed in terms of relevance and impact.

During this research, several resources where create and made available to the community. These resources include a new proposed methodology focused on health data science, several techniques for exploiting existing health datasets, new features and advancements in data science libraries (e.g.: auto-skearn or AutoKeras), or methods for optimizing machine learning architectures.

The thesis team is very hopeful that this research will spur additional data science initiatives in the field of health, including those unrelated to children.

Keywords: Health data science; eXplainable Artificial Intelligence; NAS and architectures optimization; Data science techniques for health; Children health improvements

Contents

Dedication i

Acknowledgements iii

Resumo v

Abstract vii

List of Figures xiii

List of Tables xv

Abbreviations xvii

CHAPTER 1 - Introduction 1

1.1 Thesis Publications 3

1.2 Thesis Partnerships 4

CHAPTER 2 - Literature Review 5

2.1 Data Science in Healthcare 7

2.2 Data Science in Children Healthcare 17

2.3 Novelties Presented by the Thesis 23

CHAPTER 3 - Methodology(ies) 25

3.1 CRISP-DM 25

3.2 Thesis Framework 29

3.3 Optimization Algorithms Benchmark 35

CHAPTER 4 - Framework Methods and Tools 39

4.1 Architecture Optimization 39

4.2 Explanation 47

4.3 Research Results 59

CHAPTER 5 - Derivative Works (secondary outcomes) 61

5.1 Neural Architecture Search for (1D) CNNs 62

5.2 Developments and Implementations 88

CHAPTER 6 - Applied Data Science: Child's Target Height Prediction 89

6.1 Introduction 90

6.2 Related Work 92

6.3 Business Understanding 93

6.4 Data Understanding 94

6.5 Data Preparation 100

6.6 Modelling 102

6.7 Architecture Optimization 105

6.8 Evaluation 108
6.9 Explanation 111
6.10 Deployment 113
6.11 Research Results 114
6.12 Conclusions 115
CHAPTER 7 - Applied Data Science: Length of Stay in Neonatal Intensive Care Units 117
7.1 Introduction 117
7.2 Related Work 118
7.3 Business Understanding 120
7.4 Data Understanding 122
7.5 Data Preparation 126
7.6 Modelling 128
7.7 Architecture Optimization 129
7.8 Evaluation 131
7.9 Explanation 132
7.10 Deployment 137
7.11 Research Results 137
7.12 Conclusions 140
CHAPTER 8 - Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children 143
8.1 Introduction 143
8.2 Related Work 144
8.3 Business Understanding 145
8.4 Data Understanding 147
8.5 Data Preparation 151
8.6 Modelling 153
8.7 Architecture Optimization 154
8.8 Evaluation 157
8.9 Explanation 158
8.10 Deployment 159
8.11 Research Results 160
8.12 Conclusions 161
CHAPTER 9 - Conclusions & Future Work 163
9.1 Conclusions 163
9.2 Future Work 165
References 167

Appendix 188

List of Figures

Figure 2.1 Factors influencing health status of an individual 5
Figure 2.2 AI, Machine Learning & Data Science context 9
Figure 2.3 Conceptual comparison of “traditional” programming and machine learning 10
Figure 2.4 Machine Learning Segments 11
Figure 2.5 Illustration of the unsupervised method of anomaly detection 12
Figure 3.1 Phases of the Industry Standard Process for Data Mining (CRISP-DM) reference model (adapted from the original diagram) 26
Figure 3.2 Thesis proposed framework. Three new phases (highlighted in the orange rectangle) were added to the base methodology, CRISP-DM 30
Figure 3.3 Algorithm’s comparison methodology process 36
Figure 4.1 Hyperparameters Vs Parameters in a Machine Learning model 40
Figure 4.2 Machine learning architecture illustration 41
Figure 4.3 Illustration of the random search strategy 42
Figure 4.4 Illustration of the Greedy search strategy 43
Figure 4.5 Illustration of the Bayesian search strategy 44
Figure 4.6 Illustration of the Hyperband search strategy 45
Figure 4.7 Illustration of the Genetic search strategy 46
Figure 4.8 Illustration of the Explanation concept 47
Figure 4.9 Linear regression illustrative plot and expressions 47
Figure 4.10 Decision tree algorithm illustrative explanation 49
Figure 4.11 Permutation Importance method illustration 50
Figure 4.12 Individual Conditional Expectation illustration, with a linear and non-linear example (respectively X_1 on the left and X_2 on the right) 51
Figure 4.13 Examples of Partial dependence plots using 1D (left) and 2D (right), respectively for analysing one or two features 52
Figure 4.14 SHAP prediction instance example 54
Figure 4.15 Grad-CAM method illustration 57
Figure 5.1 Illustration of a CNN architecture 64
Figure 5.2 An example of a CNN architecture definition (including hyperparameters) 66
Figure 5.3 Illustration of the Neural Architecture Search (NAS) concept 67
Figure 5.4 1D data example (heartbeat signal) 71
Figure 5.5 Illustration of the NAS strategies applied in experiments 73
Figure 5.6 Optimized 1D CNN architecture produced by the NAS process. 75
Figure 5.7 Experiment’s performance for Approach A1 (Greedy) accuracy (in %) 76
Figure 5.8 Experiment’s performance for Approach A2 (Bayesian) accuracy (in %) 77
Figure 5.9 Experiment’s performance for Approach A3 (Hyperband) accuracy (in %) 78
Figure 5.10 Experiment’s performance for Approach A4 (Random) accuracy (in %) 78
Figure 5.11 Experiment’s performance for Approach B (Genetic) accuracy (in %) 79
Figure 6.1 Distribution of heights, including for mothers, fathers, child’s, and both male and female child’s 96
Figure 6.2 Relation between father height and child height (regardless of gender). 96
Figure 6.3 Relation between father height and male child height 97
Figure 6.4 Relation between father height and female child height. 97
Figure 6.5 Regression analysis considering (both) parents and female child’s height 99
Figure 6.6 Regression analysis based on height of the female child’s and the father’s 99
Figure 6.7 Accuracy improvement in the modelling and optimization phases 107

Figure 6.8 LightGBM (red) and Mid-Parental (green) prediction outputs. 109
Figure 6.9 LightGBM (red), Mid-Parental (green) model's prediction and the child's actual height (blue) 109
Figure 6.10 Effect of gender on predicting a child's height (dark orchid for the female gender and cadet blue for the male gender) 111
Figure 6.11 Parent's influence on child´s height, considering the child gender 112
Figure 6.12 Influence of parent's height on each female child's height (red for positive influence and blue for negative influence). 113
Figure 6.13 Interface for Child's Target Height Prediction tool 114
Figure 7.1 Example of plots created in the data exploration process..... 125
Figure 7.2 LOS distribution among all dataset (highlight for the incidence of outliers) 126
Figure 7.3 Features impact on LOS prediction (record with HAI) 133
Figure 7.4 Features impact on LOS prediction (record with HAI); another perspective..... 134
Figure 7.5 Features impact on LOS prediction (record without HAI)..... 134
Figure 7.6 Factors influence and impact on LOS 135
Figure 7.7 Detail analysis for the three most relevant features for LOS: a) infections number cases; b) birth weight; c) gestational age..... 136
Figure 7.8 Impact on LOS by number of episodes of HAI and LOS total duration 138
Figure 7.9 Impact of each BW category on LOS 139
Figure 7.10 Impact of GA on LOS..... 140
Figure 8.1 An example of an extracted XML file after anonymization..... 148
Figure 8.2 Distribution of initial ECG record extraction, by age classes..... 148
Figure 8.3 Distribution of initial ECG record extraction, between 0 and 15 years of age..... 149
Figure 8.4 Example of a signal (the ECG lead) extract from a collected ECG 150
Figure 8.5 BMI statistical analysis (distribution and counting) 151
Figure 8.6 Example of ECG signal filtered with R-Peak and "complete" beats identification..... 153
Figure 8.7 Beats signal mean and standard deviation bands, stratified by BMI classes..... 153
Figure 8.8 NAS process outcome, the best NN architecture for the research problem 155
Figure 8.9 Most relevant moments for heartbeat classification using Grad-CAM++: A) Normal B) Obese 159
Figure 8.10 Grad-CAM analysis of both classes..... 160
Figure 8.11 Grad-CAM results with the representation of the two heartbeat classes shapes..... 161

List of Tables

Table 4-1 AI explainability techniques resume comparation 58
Table 5-1 Summary table for Approach A1 - Greedy accuracy (in %) 76
Table 5-2 Summary table for Approach A2 - Bayesian accuracy (in %) 77
Table 5-3 Summary table for Approach A3 – Hyperband accuracy (in %) 77
Table 5-4 Summary table for Approach A4 – Random accuracy (in %) 78
Table 5-5 Summary table for Approach B - Genetic accuracy (in %) 79
Table 5-6 Experiments measures: running time 79
Table 5-7 Experiment's measure: CPU consumption 81
Table 5-8 Experiment's measure: GPU consumptions 82
Table 5-9 Experiments Testing and Measurement overview 83
Table 6-1 Models found in the literature research (the challenge reference models) 94
Table 6-2 Height's correlation considering different perspectives, namely: child's considering gender and regardless of gender; father's and mother's height quarters 98
Table 6-3 Dataset resume 100
Table 6-4 Model's training dataset 102
Table 6-5 Testing dataset for model's performance evaluation 102
Table 6-6 Machine learning algorithms performance with default hyperparameters 104
Table 6-7 Machine learning algorithms performance after applying MaxAbsScaler scaling technique 105
Table 6-8 Machine learning algorithms performance after hyperparameters tunning 106
Table 6-9 Resources consumption for the machine learning algorithms optimization 108
Table 7-1 Models performance for LOS prediction, using default hyperparameters 128
Table 7-2 Architecture optimization best results for regression algorithms and NNs 130
Table 7-3 LGBM regressor algorithm performance excluding outliers' records 131
Table 8-1 Generation XXI cohort 146
Table 8-2 BMI classification considering BMI measure 150
Table 8-3 Confusion matrix performed on validation dataset 156
Table 8-4 Dataset's classification accuracy 156

Abbreviations

AI	Artificial intelligence
ALE	Accumulated local effects
ANN	Artificial neural network
BMI	Body mass index
BW	Birth weight
CNN	Convolutional neural network
CNN 1D	Convolutional neural network, based on 1D data input and 1D convolutions
CRISP-DM	CRoss Industry Standard Process for Data Mining
CV	Cardiovascular
CVD	Cardiovascular disease
DM	Data mining
DL	Deep learning
DT	Decision tree
ECG	Electrocardiogram
EDA	Electrodermal activity
EEG	Electroencephalography
EHR	Electronic health record
EMG	Electromyography
GA	Gestational age
GDP	Gross domestic product
HAI	Hospital-acquired infections
ICE	Individual conditional expectation
IRN	Instituto dos Registos e do Notariado
ISCTE	Iscte - Instituto Universitário de Lisboa
ISCTE-IUL	Iscte - Instituto Universitário de Lisboa
LOS	Length of stay (in a hospital)
MAE	Mean Absolute Error
ML	Machine learning
MSE	Mean Squared Error
NAS	Neural architecture search
NICU	Neonatal intensive care unit
NLP	Natural language processing

NN	Neural network. Referring to artificial neural network
OLS	Ordinary Least Squares
PDP	Partial dependence plots
PPG	Photoplethysmogram
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
SotA	State-of-the-Art
SVM	Support vector machine
VLBW	Very-low birth weight
WHO	World Health Organization
XAI	eXplainable artificial intelligence
1D	One dimensional
1D CNN	One Dimensional Convolutional Neural Networks. Also, could be referred as CNN 1D
1D NAS	One Dimensional Neural Architecture Search

CHAPTER 1 - Introduction

This PhD thesis emerged from a set of initial ideas and facts, namely: a) the society has a heavy expense on healthcare expenditures, which represent a significant part of the gross domestic product (GDP) for developed countries, and the value is increasing towards economic unsustainability; b) the healthcare services are under pressure due to a high demand and lack of resources. The recent Covid-19 pandemic has further deteriorated the situation; c) healthcare is becoming progressively digitalized and health records are being increasingly enriched with relevant data, progressing towards the creation of comprehensive electronic health records (EHR); d) the increasing adoption of wearables (e.g: smartbands and smartwatches) in daily activities and routines capture relevant data, which today is used only for "entertaining" purposes or simply as gadgets, but could be explored in other relevant ways; e) the technological progress on big data and the ability of process ever-increasing data volumes; f) a growing capacity and ability to explore data and extract insights, knowledge and value with the support of artificial intelligence, machine learning and data science techniques.

The combination of these facts underpins the opportunity to develop new processes, techniques, and methods to benefit our society, and specifically with the improvement of healthcare.

Nowadays, is possible to find several cases of successful artificial intelligence and data science projects running in live scenarios in diverse business areas, such as, fraud and risk detection on banking, virtual assistance on marketing and e-commerce, targeted advertising in search and products recommendations, and so on. Still, in the healthcare field it's still difficult do find successful production projects with significant adoption. The most relevant data science projects in healthcare are limited silos related with medical imaging analysis, such as X-Ray or skin moles image inspection, or speech to text functionalities, for instance converting medical notes to records.

In general, the successful data science projects, regardless of the business areas, were typically based in scenarios with abundant digitalized data, developed from the scratch, based in a planned and structured data collection process oriented to data science, and followed by an organized and well defined project. Perhaps the most popular examples are the Netflix recommendation system for media contents, the DeepFace project developed by Facebook which identifies which persons are in a picture, or the autonomous driving provide by Tesla still on a premature phase.

All these projects have in common enormous investments and massive data collection processes, tailored to the projects needs and goals.

This thesis aimed to explore (existing) real data sources encompassing personal data, which were previously collected without considering any data science or machine learning projects and/or goals. The thesis goal was to understand whether it was possible to extract knowledge and value from existing data that could contribute to the health and well-being of our society, using this type of data sources. Considering the extensive scope of the topic it was decided to concentrate our research in the young's segment of the population, approaching the topic from early ages.

The thesis research question can be defined as: is it possible to contribute for the health and well-being of our society exploring pre-existing datasets (collected with different purposes) with data science techniques and tools in conjunction with artificial intelligence algorithms, in a way to extract knowledge and value, in this particular case with a focus on children's health?

Data included in the research could be gathered from any (already) existing records, on diverse contexts such as medical consultations, simple biomedical records, the "traditional" clinical follow-up of child's development, the use of gadgets and wearables, or any other relevant sources of data. For instance, non-health related projects, which include data that might contribute in some way to the society health and well-being. The mandatory requirement was to have the research based in data collected from real-world cases, which were collected and/or used for a purpose other than data science.

The greatest contribution of this thesis can be resumed as the extraction of knowledge and value that was hidden until now in the data (in an immensity of data that was already been produced), contributing to the improvement of the health and well-being of children in our society with new knowledge and/or answers to real medical questions and/or making available new digital tools to explore existing health data. The thesis also introduces techniques and methodologies which may be used in other health challenges, which provides the chance to develop new initiatives, researches and projects which can continue to contribute to our society's evolution and knowledge on healthcare.

The fact that this research is supported by data collected in real-world contexts, shared with this thesis research by actual health projects (e.g.: Geração XXI, Porto, Portugal), rather than in laboratory and/or research environments (controlled environments), represents an additional challenge. The challenge encompasses the management of incomplete, irregular, incoherent, flawed data (human and/or technical), non-goal driven data, among other pitfalls inherent to a real scenario. In our perspective also introduces greater realism into the achieved results as well as it increases the probability of success when applied in other real-life scenarios.

In a way to achieve the research goal, namely the extraction of value and knowledge which can benefit our society, several digital tools were applied, such as data science processes, which included diverse state-of-the-art (SotA) methods, techniques and technologies, ranging from data exploration and visualization, validation, transformation, or feature engineering, to the latest innovations in the area of neural networks, deep learning, neural architecture search, or explainable artificial intelligence. Medical teams collaborated to develop and confirm the knowledge extraction and conclusions.

The thesis is supported on scientific certified methodologies, which were implemented, developed and/or extended with new approaches, techniques, and methods. These novelties include, for instance, the proposal for a new framework which was critical to achieve the proposed goals, neural architecture search for 1D CNNs, comparation of 1D neural architecture search approaches, or the application of XAI for knowledge extraction in healthcare contexts.

The research work, besides being able to prove the potential for extract hidden health data value for existing datasets (e.g.: cardiovascular changes due to obesity in children, or neonatal infection factors impact), also delivers answers to real clinical scientific questions (provided by the clinical teams involved) and made available developments of techniques and tools which can be used in future clinical contexts.

1.1 Thesis Publications

In the thesis context, the following research projects were published as papers:

- J. Rala Cordeiro, O. Postolache, J. C. Ferreira, “Child’s target height prediction evolution”, Applied Sciences, 9(24) 5447, 2019 [1]
- J. Rala Cordeiro, O. Postolache, “Length of stay analysis at neonatal care units with data science-preliminary results", 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA), IEEE, pp1-6, 2021 [2]
- J. Rala Cordeiro, O. Postolache, A. Raimundo, P. Sebastião, “Neural Architecture Search for 1D CNNs—Different Approaches Tests and Measurements”, Sensors, 21(23), 7990, 2021 [3]
- J. Rala Cordeiro, O. Postolache, “Big data storage for a health predictive system”, 2018 International Symposium in Sensing and Instrumentation in IoT Era (ISSI), pp1-6, 2018 [4]
- O. Postolache, L. Teixeira, J. Rala Cordeiro, L. Lima, P. Arriaga, M. Rodrigues, P. Girão, “Tailored virtual reality for smart physiotherapy”, 2019 11th International Symposium on Advanced Topics in Electrical Engineering (ATEE), IEEE, pp. 1-6, 2019 [5]

Highlight for the fact that two of the research works are published in journal with Q1 classification.

The following articles are currently being prepared for submission to relevant scientific journals:

- Length of Stay at Neonatal Care Units Analysis. A New Approach with Data Science
- Analysis of Obesity and Cardiovascular Changes Relation with Data Science

1.2 Thesis Partnerships

In the context of the thesis research, several Portuguese and European entities, health and non-health related, were contacted in order to present the research challenge. Three organisations have emerged with interesting topics, with pertinent unanswered questions, and the capacity to grant access to the appropriated datasets access.

Two partnerships were established, namely with Direção-Geral da Saúde which led to the research presented in CHAPTER 7 - Applied Data Science: Length of Stay in Neonatal Intensive Care Units, and with the Geração XXI project, from Faculdade de Medicina da Universidade do Porto whose collaboration is documented in CHAPTER 8 - Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children.

In both cases, the collaboration was based on more than just data access; it also involved knowledge sharing and the participation of clinical professionals in the research process, which was, in our opinion, of utmost importance for achieving the thesis goals.

A third partnership was planned with Instituto dos Registos e do Notariado in a way to extend even further the use case presented in CHAPTER 6 - Applied Data Science: Child's Target Height Prediction, but the process of accessing to the Portuguese population data is complex and, as at the thesis' conclusion, is still in progress.

CHAPTER 2 - Literature Review

According with some studies, there are several factors that influence a person's health status. Dr John Kitzhaber, a healthcare policy expert, has presented at the World Healthcare Congress 2014 the following distribution: lifestyle (40%), biology/genetics (30%), social context (15%), healthcare (10%) and environment factors (5%) [6].

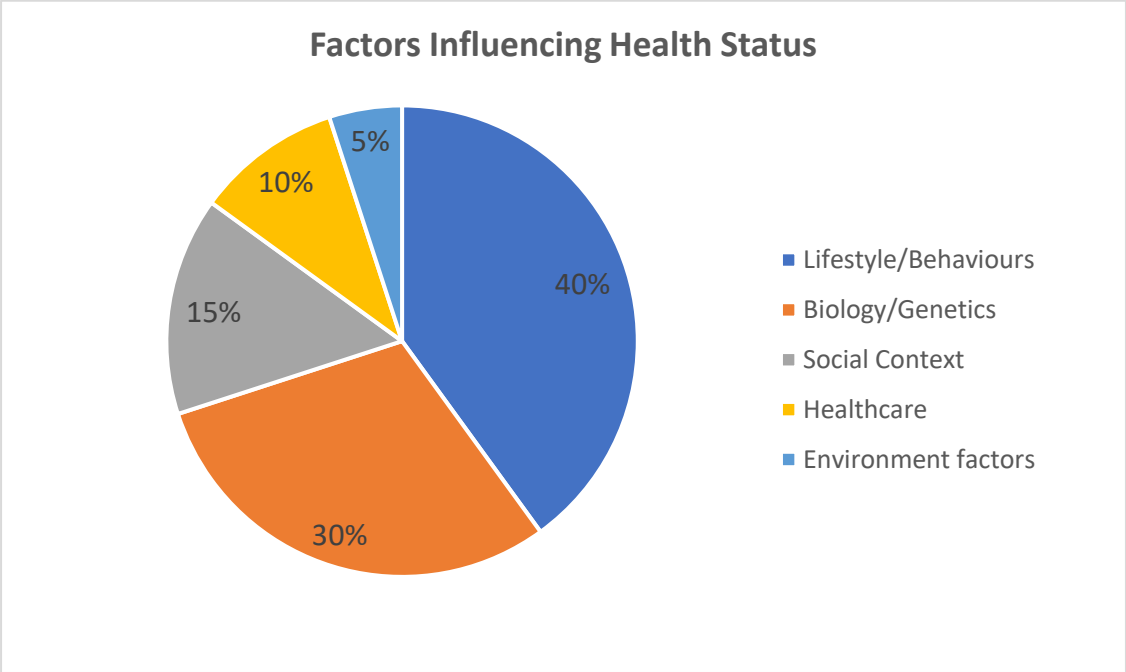


Figure 2.1 Factors influencing health status of an individual

Worldwide the amount invested in healthcare is a heavy expense, reaching as much as 16% of the GDP in United States and 9,4% in Europe with costs continuing to rise toward unsustainability. According to the previous graph these costs represent only a 10% influence on the individual's health status. Given these facts, it is essential to start exploring the other factors in order to improve the health status of the population while moving towards a sustainable model [7] [8] [9] [10].

Although the considerable investment in healthcare, the population ageing, the growth of chronic diseases (also due to ageing), the development of new clinical tests, procedures and treatments, the medicalisation of risk factors or simply the active screening of people health conditions, demand the use of more resources, both human and financial. All these factors lead to high expenditures and workforce shortages and require alternative models in order to assure sustainable, while maintaining a valuable and high quality service. The situation has become even more stressed with the challenges posed by the Covid-19 pandemic [11] [12] [13].

Healthcare units started to adopted information technologies on the 1960s, but only in the last two decades the creation of patient records, databases and the adoption of electronic health records became widespread. The digitalization of patient data has opened several opportunities, such as the ability to access quickly and from anywhere to relevant patient data, reminders of possible medication conflicts, possibility of achieving a more personalized and preventive medicine, or the opportunity to exploit huge amounts of data to better support clinical decision and operations. Nevertheless, society still lacks examples of personalized and preventive medicine, or the achievement of relevant discoveries based on data discover/mining in the health sector [14] [15].

Simultaneously with the implementation of EHR we have also witnessing a significant adoption of wearables, mobile applications and a combination of both, which collects a wide range of information about citizen activities at home, playing sports, working or in any other daily activities. These devices range from smartwatches, finger rings, hand belts, or smart bracelets to more complex medical devices such as oximeters, electronic stethoscopes or smart clothes. Nevertheless, the interest and the potential of having continuous collection and availability of personal clinical data, for instance, for taking a decision with more complete information in an emergency scenario, at the moment, these devices are being exclusively (or almost exclusively) used as gadgets, or for diagnostics of very particular clinical situations. There is a huge potential in the use of this technologies to improve our society health [16] [17] [18].

The increase of generated data was also followed by an enormous growing capacity of processing big data and a proliferation of tools, methods and techniques which allow to explore and better understand data, allowing to extract knowledge and value. Big data processing, cloud processing, data science, machine learning and artificial intelligence are some of the concepts supporting these novelties, with a continuous evolution and recurrent discoveries every day. What some refer as “a revolution” is transforming our society and our industries all over the world with Andrew Ng, a Stanford Professor and a relevant contributor to AI research, stating that “Artificial Intelligence is the new electricity”, which is a daring statement [19] [20] [21] [22] [23].

In the past years we have seen the application of artificial intelligence and machine learning in several uses cases, with remarkable advances for instance in audio and image recognition, in recommender systems, in gameplaying (such as chess and Alpha Go) or in natural language processing, currently one of the most dynamic fields in the AI area. This evolution can be found in all business areas, such as marketing, finances, robotics, automobiles, education, space exploration, energy, agriculture, manufacturing, communications and off course in healthcare [24] [25] [26].

Some popular examples of the application of these technologies are the image processing made by Domino's evaluating pizzas quality before deliver them to the client, the recommendations for new sound products, namely singers, bands and music which Spotify offers to their listeners, the help which Alibaba, Amazon or Netflix provides when assisting customers searching for new products, the estimated travel time for Uber cars, the personal home assistants such as Alexa or Siri which are increasingly "invading" our homes, or Watson by IBM that won the famous Jeopardy, a TV questions show. Many other examples across industries could have been added to this list.

The development of these systems had all in common massive amounts of data and large investments, which included large teams of experts and projects designed according to data science and machine learning projects guidelines. One of the most critical phases in these types of projects is the data collection/gathering which in these cases was carefully design and performed considering the future data science projects [26] [27] [28].

Particularly in the healthcare area there are sometimes the idea, and a kind of "vibe", that artificial intelligence and machine learning can be found everywhere, and that it is delivering benefits in many cases. When observing it with more detail, currently, we can observe a significant interest from the research community, abundant published research work and some experimental projects, but when reviewing real applications in real scenarios the success examples are still very scarce. The vast majority of the AI projects in the healthcare are still focused on image or natural language processing and more recently in the Covid-19 pandemic, in tasks such as image analysis, extraction of text or entities, contact tracking/spreading or drugs discovery. Perhaps the most popular real scenarios of artificial intelligence and machine learning adoption in the health sector are still the possibility of book appointments and pay services with the WeChat messaging platform, the Tencent Miying AI medical imaging and diagnosing platform which supports doctors in the medical image analysis (MRI and X-Ray), or the IBM's "Watson for Oncology "assisting doctors in cancer treatments [27] [29] [30] [31] [32].

2.1 Data Science in Healthcare

In the last few years there has been a considerable and rapid progress in data science, machine learning, and artificial intelligence in general. The advancements in data collection, access to computational resources, processing power, and data science talent are enabling this evolution. The healthcare industry is no exception and is gradually introducing these novelties [33].

Before going forward with the literature review of these concepts in the context of healthcare, it is crucial to understand how they relate with each other. Machine learning is a way to implement artificial intelligence and also considered a sub-category or a branch of AI. Briefly, AI was coined by John McCarth in mid-1950s and can be described as a branch of computer science where the combination of science and engineering will allow the machine to become intelligent for human welfare [34] [35].

The idea behind ML is the ability to learn from the past data/experience in a way to predict future, somehow how humans learn. In other words, is the machine's ability to learn autonomously based on experiences, observations and pattern recognition, without being explicitly programmed. The major advantage is that computers can handle much more data with much more complexity. Humans may still be able to make better predictions in some cases, but they can't compete on complexity and scale [34].

Machine learning is interdisciplinary employing techniques from a variety of fields such as computer science, mathematics, statistics, and artificial intelligence. The goal of machine learning is to develop a generalisation model, that is, a model thar can learn from a set of data (training data) in order to forecast using never-before-seen data.

Like statistics, machine learning relies on samples to infer and produce generalisations. Statistics typically focused on descriptive data (although it can also be used for extrapolation of predictions), while machine learning focuses on predictions through the application of algorithms in a way to extract insights and patterns from data.

Data science encompasses machine learning and other analytical processes, statistics and related branches of mathematics. Data science consists in the study of data, whether structured or unstructured, for extracting meaning and/or insights, with the aid of several methods, tools, algorithms and systems. That knowledge allows to "tell stories", usually using visualizations (pictures/graphics), which can benefit a diversity of entities such as companies, governments or academic institutions.

Figure 2.2 AI, Machine Learning & Data Science context - presents an illustration of all this concepts relation.

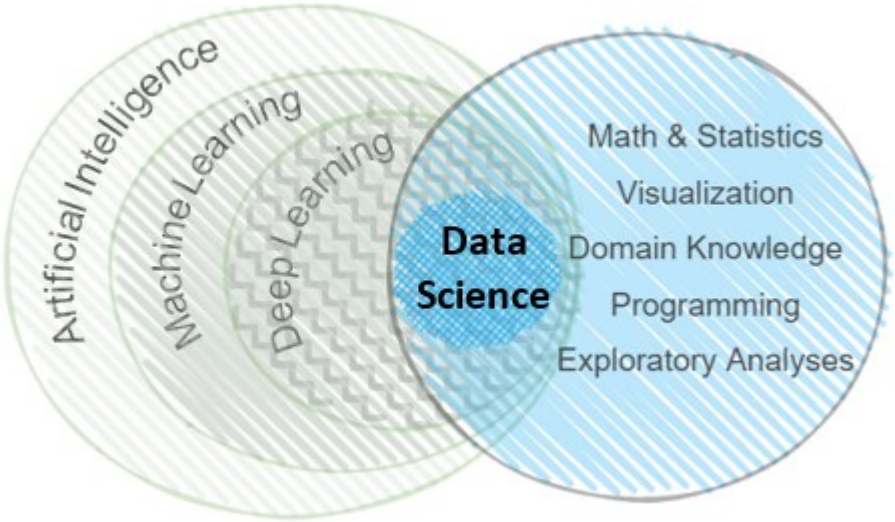


Figure 2.2 AI, Machine Learning & Data Science context

Clinical practise, medical research, and biomedical research are just a few of the areas where data science and machine learning are now used and/or increasing. Applications in these areas can be found in fields as diverse as disease diagnosis, drug discovery, treatment selection, automated assistant surgery, patient monitoring, prediction of chemical toxicity, biomarker discovery, gene function annotation, simulation of molecular dynamics, literature mining, and many more [36] [37] [38] [39].

The introduction of rule-based systems and clinical decision support systems roughly seventy years ago signalled the start of the application of "artificial intelligence" in healthcare. Back then, the term AI could be used to describe these systems, however, today the term has a different connotation. Methods for interpreting electrocardiograms (ECGs), diagnose diseases, assisting with treatment decisions, and creation of hypotheses when diagnosing complex situations arrived a few years later. These methods needed programmatic definition of all interactions, rules, and incorporation of background knowledge, which is a highly challenging, if not impossible endeavour [40] [41] [42].

The publication “ABCDE - an evolving concept in the early detection of melanoma” provides an example of a rule-based approach. Here, rules were developed based on the following standards: A describes the suspected tumour’s geometric asymmetry, B describes its irregular borders, C describes its colour mixture, D examines the lesion's diameter, and E describes how the lesion is spreading or growing [43].

Data science techniques nowadays, and in particular machine learning techniques, take a different approach by leveraging the algorithms' potential to incorporate knowledge that, up until recently, required specialised programming but is now inferred by the algorithms [44] [45].

With machine learning there is no explicit programming of rules, behaviours, or other relations. Instead, the computer program learns how to implement the model using past data. Learning is a process made by searching meaningful patterns in data, which is essentially "connecting" a number of inputs to the output(s) [45].

The differences between the machine learning approach and the previous programming methods can be illustrated by the Figure 2.3.

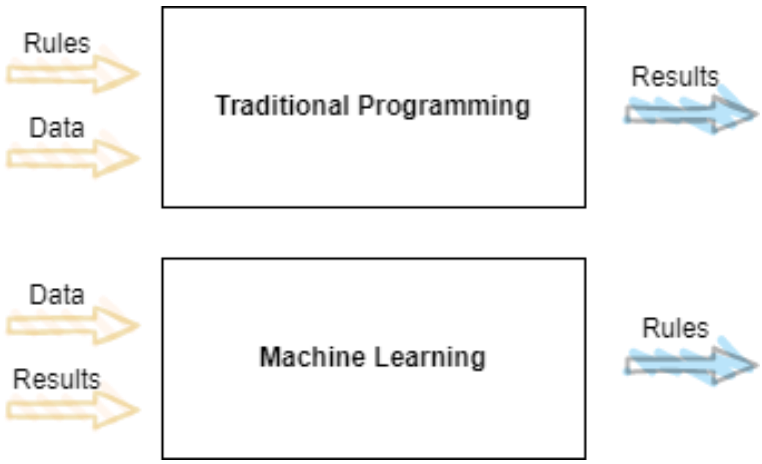


Figure 2.3 Conceptual comparison of "traditional" programming and machine learning

In ML there is no need to explicitly build a model, which is the responsibility for the ML algorithms. In this case, the programmer main concerns are the choice of algorithms and the model adjustments that will increase performance. Also, ML allows to process large amounts of inputs parameters, as long it has enough resources (processing and memory capacity). That would be very challenging for a human to implement, as it would need to take all the details into account.

ML is commonly divided into supervised and unsupervised, but it is not uncommon to find literature which can introduce reinforcement learning as another category [34].

The supervised segment entails learning to predict outputs, which can take the form of labels, categories and continuous values, where both input and the corresponding output(s) are known. Both inputs and outputs are used during the model training process. The designation of supervised learning is based on the fact that the learning process is aware of the desired outcomes.

Problems in supervised learning can be divided into classification and regression problems. The goal of a classification problem is to group elements into a class or category. For instance: positive, negative; sick, healthy; tumour, unknow mass, no mass; treatments options. It is a method that enables the prediction of a group for each data instance. A regression problem, in the other hand, learns a function by considering historical data in order to predict a number. For instance, predicting the length of stay in a hospital unit, a patient scoring risk, or the expected number of patients in a hospital emergency room.

Supervised machine learning models can also be broken down into discriminative models and generative models. Discriminative models are able to recognise the “boundaries” between different classes or categories and produce the labels that correspond as an output, while generative models produce new instances of a given class. For instance, a discriminative model can categorise patients into different groups, labelling patients according to a particular risk (e.g.: breast cancer risk), while a generative model can personalize a drug for a particular patient considering his characteristics.

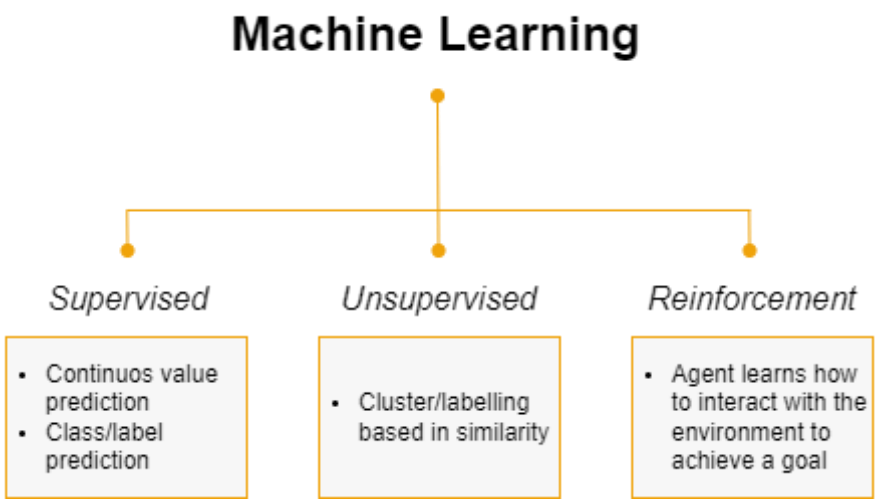


Figure 2.4 Machine Learning Segments

Unsupervised learning groups the dataset's instances according to similarities. Unlike supervising learning, the training process doesn't involve labels. This ML segment includes methods such as clustering, dimensionality reduction, anomaly detection, and recommendation systems.

The clustering methods divides data instances into groups of instances with similar characteristics such as patients' reactions to a particular drug. Depending on the knowledge around the challenge, there are various ways to process the clustering. How to measure similarity and whether the number of clusters is known are important pieces of information to consider when choosing a technique.

Regarding anomaly detection, it is essential to start with the definition of an anomaly. An anomaly is characterized by something that deviates from the expectable. E.g.: a doctor who charges 1000€ per hour or an alteration in a patient heartbeat. On this approach, it is common to have unbalanced datasets where the anomaly class is represented by further fewer items than the "regular" class, and algorithms must be suitable for this condition. Also, anomalies frequently occur in a variety of formats and shapes, and new kinds of anomalies can appear at any time. Advantages of anomalies detection include not needing to define types of anomalies in advance, as a supervised method would require.

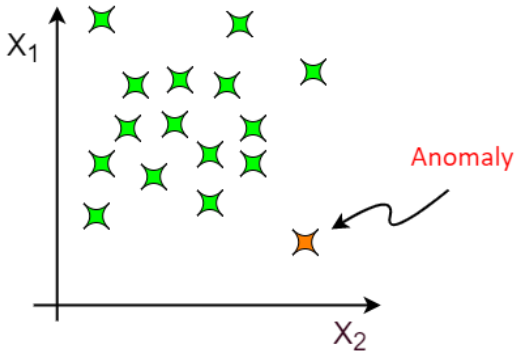


Figure 2.5 Illustration of the unsupervised method of anomaly detection

Given that the process also doesn't require the use of labels, reinforcement learning can be viewed as a subset of unsupervised learning. This methods studies how a software agent behaves and takes actions in a virtual environment, which often simulates a real-world scenario.

It's a goal oriented method, which learns how to attain a complex goal. The method will accurately define the best path for the agent to take and will discover the best possible actions for any given state through a process of trial and error, enabling the goal of reaching a certain point. The agent receives a reward (or, eventually, a punishment) for every action, and the goal is to discover how to maximise the reward over the long term by utilising and exploring the knowledge contained in the data.

These methods are frequently used to enable software to play games, enhance robot performance (by helping them adapt to various environments), or to aid in vehicle navigation, but can also be applied in healthcare domains. For instance, the chronic diseases treatment can be seen as sequential decision-making problem, and these methods can assist on the optimization of drugs timing and dosage, evaluating the treatment outcomes to promote the patient's long-term benefits [46].

The growth of annotated clinical data and the simplification/facilitation of access to data science (e.g.: open source libraries, computational resources costs, available expertise) have powered in particular diagnose tasks, typically supported by supervised learning approaches. The most successful area of medical AI at the moment is undoubtedly automated medical-image diagnosis, which benefits a variety of medical specialties that rely in image-based diagnoses (e.g: dermatology, ophthalmology, radiology, or pathology) [33] [47].

The success of the imaging diagnostics is directly related to the emergence of deep learning over the last decade. Convolutional neural networks, a deep learning architecture, in particular have brought significant improvements in the medical-image analysis domain, and more broadly in the field of computer vision. Before that, multidisciplinary teams of physicians and developers had to manually create/extract (useful) features from the image training set, which frequently led to time-consuming and imperfect processes. Convolutional neural networks (CNN s) have made it possible to learn useful features from the image training set automatically [48].

For instance, the popular image processing models developed by the Google Brain AI research team were able to detect diabetes sequelae in the eyes at an early stage. These algorithms scan retinal images (retinal screening) in search of aneurisms that may indicate diabetic retinopathy and those helping in the prevention of this disease that can cause blindness. The study's results were interesting, but they also highlight to the need for further research [49].

Kuochen Zhou et al. proposed a deep learning process to analyse mammographic images for the purpose of early breast cancer detection. The model was based on the "scientific" datasets mini-MIAS (322 mammographic images) and INbreast (410 images), outperforming the previous researches in the field. In this case it was also address the ability to use transfer learning with pre-trained network segmenting (Deeplab v3), the main goal of the study [50].

One of the many other applications of image processing in healthcare is dermatology. Noel Codell develop an application that can identify melanomas (skin cancer) using the dataset made available by the International Skin Imaging Collaboration (900 images). Previous studies were focused on surpassing state-of-the-art performance, but no clinical implementation was made [51].

There are many other research cases in the field of health image processing. Research works generally use "scientific" datasets, frequently with few images, collected from a single or a very limited number of hospital units, and are concentrated on the model's performance and in overcoming the SotA performance. Limitations of general studies include data imbalance, sample size, sample quality, lack of generalisation, and ambiguity regarding how models produce results, among others issues [52] [53] [54] [55] [56] [57] [58].

Large-scale and high-quality training data are crucial for machine learning supervised algorithms. The lack of annotated images and the fact that creating these kinds of datasets takes time, requires expertise and is therefore expensive, are frequently cited as obstacles. Text mining on pre-existing reports can be used to generate annotations, making the process quicker and less expensive, but it is still important to assess the quality of the generated annotations.

Despite the challenges in the field of image-based diagnostics, some systems have already received approval from medical regulatory authorities (like the U.S. Food and Drug Administration) and are fully functional in real environments. The specialty with the greatest representation—more than 70% of the approved systems—is radiology. Examples of approved systems include ProFound AI, which uses digital breast tomosynthesis analyses to detect breast cancer, OsteoDetect, which uses X-rays taken of adult patients to find wrist fractures, and HealthPNX, which uses X-rays of the chest to detect pneumothorax [59].

Text processing, which includes methods like named entity recognition (NER), relation/event extraction, and information retrieval, is another expanding area in the health field [60].

Examples of NER application can be found, for instance, in the biomedical literature work of Robert Leaman and Zhiyong Lu. The authors created the TaggerOne system, which successfully extracted disease and chemical mentions, based on the NCBI Disease corpus and BioCreative V Chemical Disease Relation task corpus. Performance is described in the research discussion as "the highest of which we are aware" [61].

Another example is a study on drug name recognition that used texts collected from DDIExtraction 2013, a dataset frequently used in this kind of challenge. The study's findings were summarised by Shengyu Liu et al. as follows: "…our best system achieves much better performance. However, it is not good enough for medical information extraction systems and applications, e.g., drug–drug interaction extraction, clinical decision support and drug development" [62].

This dataset (DDIExtraction) continues to be used in similar challenges, with different approaches, as it is presented in a very recent article (November 2022) titled "Extracting drug–drug interactions from no-blinding texts using key semantic sentences and GHM loss". In this case, the research was on the interactions between drugs, and drug name recognition was required as "intermedium task" [63].

The paper ”Improving chemical disease relation extraction with rich features and weakly labelled data" falls under the category of relation/event extraction. It describes a method based on support vector machines (SVM) for determining relationships between drugs and diseases with the aim of discovering new treatments. This technique has demonstrated SotA performance in disease-chemical relation extraction. Fei Li et al. assessed relationships in biomedical texts as well, but in this case, they extracted adverse drug events. This research, which is centred on the relationship between drug and disease entities, was able to create a useful model based on neural networks that performs competitively with SotA systems [64] [65].

Drug discovery and protein relations, although they appear to be related to the previous field, required completely different approaches. The application of data science in this field also seems to be leading to some very interesting innovations.

The article "Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery" provides an overview of the history of AI's application in the field of drug discovery from the start of the second half of the 20th century to the present . This article introduces a variety of topics related to AI and data science, as well as how machine learning algorithms can help with drug discovery. It highlights how neural networks, in particular deep learning, have had a greater impact in the field with their strong capacity for pattern recognition [66].

An example of an interesting protein contact work is presented in the research paper “Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model” which uses deep leaning to model very complex protein sequence-contact relationship as well as long-range inter-contact correlations [67].

The application of data science in drug discovery is currently a "hot" topic In the scientific community. This field will undoubtedly advance in the years to come [60].

Autonomous surgery and robot-assisted surgery is another field that is being leveraged by advances in data science. The use of technology in surgery has always required the involvement of a surgeon to oversee movements due to technical limitations. However, recent innovations are empowering more autonomy for the technology [68] [69].

According to the research paper "Supervised autonomous robotic soft tissue surgery," a system called Smart Tissue Autonomous Robot performed a supervised autonomous soft tissue surgery in which an algorithm planned and executed a complex procedure on deformable soft tissue. The study concluded that the system's “autonomous procedures is superior to surgery performed by expert surgeons” [70].

Tara M. Connelly has conducted an extensive analysis on the robotic surgery research. The paper begins with a historical overview of the topic's evolution, explores the areas that are currently receiving more attention (such as urology and colorectal surgery) and discusses both the future directions and limitations on these subspecialities. The author draws the conclusion that robotic assisted has gained acceptance in many of the subspecialities. These kinds of studies are paving the way for widespread adoption in specialties like neurosurgery, urology, colorectal, cardiothoracic, orthopaedic, or maxillofacial surgery. The potential advantages robotic surgery includes a decrease in human error rates, shorter procedure times, lower costs, and improved patient access to healthcare [71].

The health data science research has also progressed in the analyses of electronic health records (EHR) as well as with records associated with daily live (e.g.: wearable devices). With the widespread adoption of EHRs and the expansion of these types of records (with more and better information), there are many exciting opportunities to research in the field of data science [72] [73].

For instance, Xiongcai Ca et al. have developed predictive systems using Bayesian networks and patient-specific administrative and laboratory data that forecast hospitalisation days, readmission days, and death simultaneously. As in many of these studies, the performance of the model was the main concern [74].

The topic of non-invasive and wearable sensors for remote health monitoring assisted by data science is develop, for instance, in the paper "Wearable Sensors for Remote Health Monitoring." The work includes examples of physiological signs monitoring and motion activity tracking enabling, for instance, elderly people to stay in their familiar surroundings rather than occupying expensive and limited healthcare facilities. Monitoring arrhythmias, heart attacks, falls, musculoskeletal or cognitive diseases, body temperature, galvanic skin response, and blood oxygen saturation are just a few of the topics covered in the study [75].

The work "How to use digital devices to detect and manage arrhythmias: an EHRA practical guide" which takes a wearables-based approach to heart rhythm monitoring was recently presented by the European Society of Cardiology in 2022. The paper discusses the use and application of some wearables approved by regulatory authorities for diagnosing cardiac abnormal findings like arrhythmias, atrial fibrillation, or ventricular tachycardia. Current medicine is starting to employ these techniques [76].

For instance, the author of this thesis is experiencing some cardiac alterations that has been hard to follow with “traditional” methods such as ECG, Holter or Echocardiogram. João has been "challenged" by his cardiologist to use a wearable "tracker" and record whenever acute episodes occur.

Although there are many other works in this field of wearables, the real application in the area of health is still in early stages. Also, it is still common to find some issues around the accuracy of the data collected through these devices [77] [78] [79].

In overview, a lot of research papers can be found when looking for artificial intelligence, machine learning and data science applications in the healthcare area. Despite the "thousands" of papers in the area, finding strong and successful real cases application is still difficult. As the author Rahul C. Deo states "… although there are thousands of papers applying machine learning algorithms to medical data, very few have contributed meaningfully to clinical care", an affirmation that was recently confirmed by Jiamin Yin et al. "Despite a great deal of research in the development and validation of healthcare AI, only few applications have been actually implemented at the frontlines of clinical practice" [80] [81].

Although data science promises to revolutionise healthcare, there are still a number of obstacles that must be overcome, including access to data, data standardisation and quality, legal and ethical concerns regarding the acquisition and use of data, effective AI system regulation for healthcare, improving the costly and time-consuming task of labelling data sources, keeping up with the technological advancements that are occurring every day, and still some mistrust of health professionals who sometimes struggle with these new methods. In this last field, explainable AI will play a significant role in producing results that are more reliable and understandable, which will contribute to increase the trust of medical professionals [82] [83] [84].

There is a lot of debate about whether AI will eventually replace health personnel. AI and data science are more likely to improve healthcare quality by reducing medical errors and the stress that routine clinical tasks put on medical staff. It is not expected that AI will decrease clinical workload, but rather increase professional's allocation to high value added tasks, for instance, by cutting down on time spent on diagnosis and allowing for more time to be spent on treatment [81] [85].

Although there is still a long way to go, these innovations are slowly gaining ground and expanding into fields that were previously thought to be the exclusive domain of human specialists.

2.2 Data Science in Children Healthcare

The conditions experienced during the first 18 years of life have a significant impact on the health and well-being of adults. Since many adult diseases are known to have their roots in infancy, childhood, adolescence, or even before birth, it is imperative to start addressing healthcare from early ages [86].

Children's healthcare and health outcomes in general have significantly improved over the past three to four decades, especially in more developed countries. The widespread use of vaccinations, along with other advancements like better nutrition and sanitation, were essential for the improvement of health [87].

The availability and growth of health-related data, along with advancements in data science and AI that increase the ability to model complex environments and interactions, promise to continue to improve children's healthcare and, by extension, the health of the population as a whole. However, currently it is possible to notice a significant discrepancy between the child health outcomes when compared with the population as a whole because of the unique characteristics of the paediatric population [88] [89].

The paediatric population poses additional challenges because of the growth-specific issues, paediatric diseases, and constant changes in maturity stages. A group of children ranging in age from 0 to 18 exhibits significant heterogeneity, which can cause models to perform well for some age groups but poorly for others (e.g., preterm neonates vs. term neonates). For instance, vital signs like heart rate and blood pressure differ significantly depending on the child's age range, and various viruses (like measles and chickenpox) behave and respond differently to treatment depending on the patient's age. These variables are in addition to those that affect the general population, such as age, gender, race, genetic predispositions, or existing diseases. When compared to adult medicine, these particularities lead to the need for datasets with greater resolution/granularity [90] [91] .

An example of this complexity is presented by Zachary Vesoulis et al., which has studied the performance of pulse oximeters in both preterm infants with black and white skin. In addition to the complexity already brought on by preterm birth, a consistent error in the SpO2 measurements was found to be a risk factor for occult hypoxemia in black infants [92].

It is also common to encounter data science challenges that are exacerbated by the fact that children are a population subset that requires safeguards. Just to name a few, these difficulties include logistical, moral, and legal constrains. For instance, in paediatric pharmacotherapy, drug administration efficacy and safety are frequently derived from adult clinical trials already in existence or defined based on fewer and smaller studies, which results in a less accurate drug characterization. [93] [94].

Research on paediatric pharmacotherapy is one area where the limitations are more noticeable and where data science can make interesting contributions. As stated by Charlotte I. S. Barker et al. in their study "Pharmacokinetic studies in children: recommendations for practise and research" determining the dosage of medications for infants and children remains difficult, and model-based approaches may be helpful in this situation [95].

In the paper "Beyond the Randomized Clinical Trial: Innovative Data Science to Close the Pediatric Evidence Gap" some potential strategies for addressing the problems about drug clinical trials involving children are presented. These strategies include using EHRs and machine learning techniques (e.g: natural-language processing) [96].

Natural language processing (NLP) besides being useful in the previous challenges can, like in adulthood healthcare, be used in a diversity of scenarios. Text mining in EHRs has been successfully applied by Kevin Bretonnel Cohen et al., who have developed an ML system that identifies potential epilepsy surgery candidates. The study makes use of the EHRs to identify patients whose seizures are "clinically intractable" and who may need neurosurgery to treat epilepsy. The study compared the performance of the system with the level of agreement between two annotators, showing that the system was able to outperform the annotators and concluding that this technique may be helpful in identifying paediatric epilepsy surgery candidates [97].

People most frequently commit suicide as a result of psychological distress or social pressure. These conditions are frequently present in children, especially those who are adolescents. The suicide is commonly signed by markers such as suicide notes, revelations in private diaries, or messages on social media. John P. Pestian et al. created an ML system based on NLP to predict the risk of adolescent suicide. They reached the conclusion that suicide assessment and prevention could be done using NLP [98].

There are other paediatric research papers using EHR and/or NLP with a variety of objectives, including extracting clinically relevant data, extracting or masking patient's identification, supporting diagnoses, and extracting labels to enable the use of records in different ML variants (e.g.: image classification). The majority of the data collected in EHRs is used primarily for clinical practise; research use can be an additional or side benefit. The multiple data formats present in these records make it difficult to capture and share medical data because there is a lack of a universal and interoperable format. Perhaps efforts to reformat and increase standardisation of EHRs can be developed in the near future, leading to records that can better contribute to knowledge creation and innovation [99] [100] [101].

Despite the interest and range of applications of data science in pharmacology, clinical trials, and EHR previous approached in this thesis section, radiology is the field that has made the most advancements in paediatrics, similar to what we observed in healthcare for adults. Some of the address radiology topics are relevant to people of all ages, while others differ based on the stage of life [102].

For example, image analysis in paediatrics has been used to track brain development and spot potential diseases earlier. In order to segment infant brain tissue images into cerebrospinal fluid, white matter, and grey matter, Wenlu Zhang et al. developed an ML model using CNNs. The authors concluded that this method outperformed similar studies using SVM, random forest, or coupled level set methods in terms of results (system performance) [103].

Skeletal development in children is evidently a relevant subject. Once more, image analysis can aid in the monitoring of this development. A skeletal maturity assessment implementation using child hands radiographs is described in the paper "Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs" The system is as accurate as a skilled radiologist and SotA studies, according to the authors [104].

Similar to the adult domain, there are numerous other cases in ML image analysis, including in segmentation of lesions and identification and classification of diseases [105] [106] [107].

Paediatrics with their specificities, presents some particular areas, namely neonatology which is an exclusive subspeciality, or some dedicated subjects in psychiatry, and neurology [108].

According to Zahra Hoodbhoy's systematic review of machine learning's contributions to paediatric healthcare, the most popular ML topics in neonatology relate around prematurity, neonatal seizures, and mortality. Disorders on the autism spectrum, hyperactivity, and mood disorders predominate in psychiatry. Neurology's most frequently discussed topics were epilepsy, cerebral palsy, and cognitive disability. China and mainly USA were the two countries that produced the most studies on these areas. Canada, Brazil, Europe, Australia and India also contribute to the topic but on a smaller scale [108].

Neonatal care includes a variety of retrospective and observational studies that cover topics like newborn diseases (such as newborn jaundice), organ dysfunction, survival rates, and risk assessments. The majority of neonatology's current tools are frequently based on static measures, such as data that is accessible at birth or within 24 hours of birth, or in hospitalization episodes. Heart rate, oxygen saturation, and blood pressure are among the vital signs monitored while the patient is in the hospital. Frequently, risk factors and complications acquired during hospitalization, like respiratory support or intravenous feeding, are not included in the tools. The use of data science is presenting alternatives to the existing techniques [109].

For instance, several ML models have been developed by Abbas Sheikhtaheri et al. to predict the risk of neonatal death in NICUs. The study used multiple layer perceptron neural networks, SVM, Bayesian networks, ensemble models, and decision trees (RF, C5.0, and CHAID). The data used in the study included information such as gestational age, birth weight, prenatal care, congenital malformations, and mother's health. The author concluded that while RF decision trees had the best specificity and precision performance, multiple layer perceptron NN had the highest accuracy, sensitivity, and area under the roc curve (AUC), a popular metric in this kind of studies [110].

Neonatal jaundice, a condition which is characterized by yellow discoloration of a newborn baby's skin and eyes, can progress to severe jaundice in 10–19% of cases, causing a paediatric emergency. The current strategies to diagnose jaundice use blood tests. The paper “Neonatal Jaundice Detection System” proposes the early diagnosis of jaundice using k-nearest neighbour and support vector regressions, avoiding an invasive process in the child. The project achieved an 85% accuracy in 40 jaundiced babies [111].

These are only a few instances of how ML might be used to improve neonatal care, but it is possible to find additional research in this area. Most continue to emphasise the importance of conducting more research to fully reap the rewards of AI/ML in these fields [112] [113].

In paediatric psychiatry, the autism spectrum disorder is the most common topic. The diagnosis of autism is the first area of concern; it is crucial to start treating the pathology as soon as possible and make sure that autistic patients receive enough support, as this is a crucial factor in their potential health outcome. A review of feature selection methods and machine learning algorithms, including the decision trees, SVMs, and neural networks (very popular algorithms in the field), are introduced in the paper "A Review of Machine Learning Methods of Feature Selection and Classification for Autism Spectrum Disorder". The paper discusses the results (performance) of various studies conducted in the area, concluding that "further research is required for accurate and efficient classification of autism spectrum disorder" [114].

On the other hand, the study "Reversal learning task in children with autism spectrum disorder: a robot-based approach" looks into how Keepon, a robotic toy, impacts children with autism spectrum disorder. It is known that children with this disorder frequently exhibit rigid and perseverative behaviours. The robot facilitates therapeutic sessions where children interact with it to enhance their cognitive, social, and response skills. The study found that children with this type of disorder appeared to enjoy tasks more when interacting with the robot than when interacting with an adult. It also highlights some restrictions (such as the population under study) and the need for additional research [115].

Similar to autism spectrum disorder, cerebral palsy requires an early diagnosis to improve the potential of patient's recovery. While there is currently no cure for cerebral palsy, supportive therapies like physical therapy and surgical procedures like orthopaedic surgery can help children with their motor skills. Although it is well known that this pathology manifests its symptoms in the first few months of life, a diagnosis is frequently made only at the age of two. It has been demonstrated that high-risk infants can be identified as having movement impairments using neuroimaging motor assessment through ML approaches. A few ML models are described in the paper "Multivariate Analysis and Machine Learning in Cerebral Palsy Research" that inspect lesions on cranial ultrasound and magnetic resonance imaging and aid in the early diagnosis of cerebral palsy [116].

In summary, paediatric health data science shares many of the same difficulties and constrains found in health data science for adults, but it also encounters some additional variables that add more complexity. These variables can come in a variety of shapes, such as the need for more granular data (age ranges), the challenge of gathering data due to communication issues or a child's unique characteristics (e.g.: calming a child to perform an ECG), ethical and legal considerations that (appropriately) protect children, childhood-specific pathologies, among others.

Following this complexity is the importance of addressing healthcare in the early years, allowing for earlier interventions that may significantly affect adult health conditions. Unfortunately, in this “conflict” between complexity and potential for early interventions, complexity prevails. Paediatric healthcare has been the subject of much less research than adult healthcare.

The field of paediatrics is also distinguished by a few subspecialties that are unique to it (such as neonatal care and development assessment), which, because of their relevance, demand a lot of resources and attention.

The evidence-based medicine in paediatrics is a concern, as it is in adult medicine, and the nature of some machine learning methods presenting "black box" results is an obstacle to the adoption of this innovations. Clear information is necessary because a lot of decisions are shared between the doctor and the patient (and/or his tutors). Model’s explainability is a crucial consideration in the decision-making process.

Despite these difficulties/characteristics, data science, machine learning, and AI offer fantastic opportunities to improve paediatric healthcare, which will ultimately improve the overall health of the population. To achieve this, it will be necessary to create effective, fair, and secure strategies for incorporating these novel technologies into paediatricians' regular daily patient care [117].

2.3 Novelties Presented by the Thesis

This thesis intends to bring some novelty to the healthcare sector. The research is a contribution for the application of artificial intelligence, machine learning and data science in healthcare which tries to approach exactly the areas that, in our opinion, lack of attention and/or research, namely:

- Our work is focused on humans at early ages, children, when the AI healthcare research is vastly conducted in humans with pre-existing and/or diagnosed diseases, which often leads to adult and elderly people.
- The current research in healthcare is dominated by projects dedicated to image or natural language processing. Our research work is dedicated to other types of data, such as raw data (e.g.: factors impacting a clinical condition, or personal characteristics which influence a clinical condition) or physiological signals.
- Often (almost exclusively) the existing research works include a dedicated phase for data collection/gathering design which are performed considering the future processing in a data science project context or uses a dataset created in these circumstances. One of the goals of this thesis was to extract insights from existing data sources, which were created with other purposes and haven't consider any data science or AI project guidelines. The goal was to verify if it was possible to extract value and knowledge which were hidden until now in existing valuable resources (existing datasets), in a way to find answers to real clinic questions and/or to benefit citizens health & well-being.
- Our research work includes the most novelty techniques and concepts in the AI area, namely the use of architecture optimization search techniques (including neural architecture search) and explainable artificial intelligence. In our opinion, the combination of this techniques delivers real advantages to healthcare, and the application of this resources is still very scarce. For instance, to the best of our knowledges our research was the first to present the combination of these techniques for signal processing using 1D Convolutional Neural Networks.

- Usually, research in healthcare consists in developing forecast or classification systems and use the results in a specific/restrict case or simply present how good was the achieved performance as a comparation with other researches. Our work also includes forecast and classification models, but mainly these models were used as a tool and not as the main goal. In the thesis research, models were developed as the basis to extract knowledge and in a way to answer to research questions, the real goal. For instance, one of the research use cases explored ECG diagnoses. A machine learning model was developed to classify the ECGs in classes, but the final goal was to comprehend the cardiovascular changes. The classification task was “just” part of the process of extracting the knowledge.
- All the performed research work was based in real scenarios with real clinical questions, providing new knowledge to the healthcare area. The goal was to benefit our society in a direct form. The use of the proposed techniques in other clinical areas is also discussed and encouraged.

CHAPTER 3 - Methodology(ies)

The PhD thesis and the associated research works were based in the CRISP-DM methodology which stands for CRoss Industry Standard Process for Data Mining. This methodology was designed in industrial processes context, but in this thesis is applied in the context of health and well-being [118] [119] [120] [121].

This methodology was not the only available choice. Before deciding what was the best base to support the research goals, a literature review was conducted about machine learning methodologies. As output, we have considered three of the most accredited ML methodologies, which were analysed in detail, namely Data Mining Methodology for Engineering (DMME); Sample, Explore, Modify, Model, and Assess (SEMMA); and CRoss Industry Standard Process for Data Mining (CRISP-DM).

CRISP-DM was developed by IBM in cooperation with other companies and is available as an open standard. This methodology assumes that data for the process has been already collected, which is one of the base requirements on this thesis. This methodology, the base process for the research works, is developed in detail in the following section 3.1 CRISP-DM [119] [121].

DMME it's an extension of the CRISP-DM methodology, presented by Steffen Huber et al., which adds to the machine learning process a data acquisition phase. It also includes concepts for process integrated data acquisition and infrastructure planning. It is commonly used in data analytics for engineering applications. SEMMA was developed by the SAS Institute and is more commonly used by SAS tools users as it is integrated into SAS tools such as Enterprise Miner, but also used in other contexts. SEMMA divides the data mining process into five stages, namely the phases which compose the methodology name. Sample: explores a statistically representation of the dataset; Explore: which applies statistical and visualization techniques to explore data; Modify: the phase in charge of select and transforming the features for the model; Model: builds the ML model; Assess: where the model's performance is analysed. The phases are performed as a cycle [120] [122].

It was decided to use as a base methodology the popular CRISP-DM methodology, as it has already been proven successful on data mining challenges, including in the clinical area, and in our point of view presented the best characteristics for the thesis particular challenges [121] [123].

3.1 CRISP-DM

CRISP-DM methodology can be resume by the following diagram, presented in Figure 3.1.

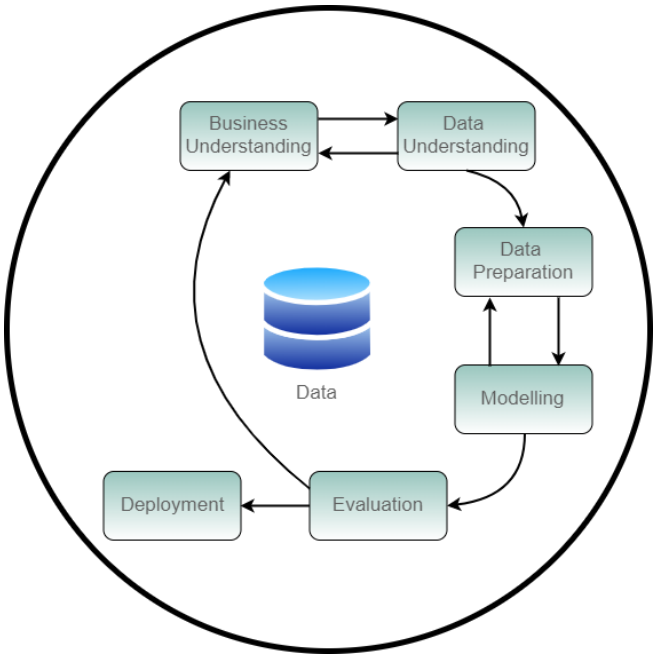


Figure 3.1 Phases of the Industry Standard Process for Data Mining (CRISP-DM) reference model (adapted from the original diagram)

This methodology consists of a hierarchical and cyclic process that breaks the data mining process into the following phases:

- Business understanding: To understand project objectives and requirements

  In this phase, a business "recognition" is made and the "problem" to be solved is defined. Getting a clear understanding of the business needs, makes it possible to convert the needs to a data mining problem definition. This definition will include the machine learning type to use (Figure 2.4 Machine Learning Segments), the criteria for model assessment and the process final goals. The process goals can be seen from 2 perspectives, namely the goal of the data mining process and the business goal.

  In a way to achieve this phase requirements the following steps should be considered: Business Understanding Overview; Determining Business Objectives; Assessing the Situation; Determining Data Mining Goals; Producing a Project Plan.

- Data understanding: Focus on becoming familiar with the dataset and identifying data quality issues using data query, inspection, and visualization

  The current stage allows for deep understanding of the data to be used. It starts with the data acquisition for the data mining process, which is then described in terms of volume and gross properties (e.g.: type of data, possible values, frequency). It is crucial to understand each attribute meaning and values in business terms.

The statistical analysis will be in charge of obtaining detail information about the attributes, as for instance mean, max, min, distribution shape, standard deviation, and correlations, which are crucial to understand relations between the different features. Let's take the length of a hospital stay as an example. The metrics used in the statistical analysis will include the longest and shortest stays, the mean and mode, and how this feature relates to others. For instance, how the length of stay relates to the gender or the infection type.

The data quality verification tasks are also included in this phase. It is necessary to perform activities such as identifying missing attributes, blank fields or invalid information, checking spelling values, duplicates, inconsistency values between features, among others.

This phase comprises the following tasks: Data Understanding Overview; Collecting Initial Data; Describing Data; Exploring Data; Verifying Data Quality.

- Data preparation: Also known as pre-processing, this phase consists of selecting and preparing data, and has as output the final dataset that will be used to train and test the model(s).

  The phase includes all necessary tasks to construct the final dataset to process and is often responsible for a huge part of the entire time dedicated to the overall data mining process (which can reach up to 90%).

  Data preparation includes activities such as data selection, correction and/or removal of errors and noise, collection of eventual additional data (internal and/or external), considerations about the use of sampling techniques, performing of features encoding, feature scaling, among other possible activities. This phase can also include the construction of “new” data, namely with resort to feature engineering which can include, for instance, building derived attributes or methods of data imputation.

  The stage is closed with data formatting for modelling, which can still include rearranging attributes, reordering records, or any special need for reformatted values (for instance for NLP challenges illegal characters or uppercase/lowercase text removal).

  The data preparation phase can be split of the following activities: Data Preparation Overview; Selecting Data; Cleaning Data; Constructing New Data; Integrating Data; Formatting Data.

- Modeling: A model which represents the existing knowledge, is built through the use of machine learning tools.

  Modelling can be essentially divided into Selection of Algorithms to be Used; Development of Training/Learning and Test Scenarios (including the partition of the base dataset); Model Creation; and Model Assessment.

On the first task, the data mining objectives are analysed in a way to define the most adequate/promissory algorithms to apply. Some techniques demand specific requirements for data format, which can lead to a new loop on the data preparation phase (e.g.: requirement of a gaussian distribution or constrain in dealing with feature in different scales).

The training and test scenarios development require several decisions such as, how to split the final dataset (result from the previous phase), plan for training and validation of data, how to evaluate model performance, or training steps (e.g.: number of iterations, the use of cross-validation, batch size).

The model creation is the effective running of model training, documenting all decisions made within the process, which include initial parameter values, optimizers, learning setting, and so on.

The modelling phase is closed with the model assessment which mainly consists of rank and evaluates model performance according to the evaluation criteria (defined in the methodology first phase) and the interpretation of results, verifying their reliability and their business value. For formally the methodology defines the following activities for this phase: Modeling Overview; Selecting Modeling Techniques; Generating a Test Design; Building the Models; Assessing the Model.

- Evaluation: Evaluate the model performance and utility in a way to decide if a new CRISP-DM iteration will be needed or if to continue to the final deployment phase.

  This phase starts with the modelling results evaluation, which implies understanding the data mining process output, and the relation with the entire process and business requirements. It is important not only to check the achievement of the data mining goal, but also to understand the business impact, namely if new knowledge or capacities were achieved. It is not uncommon to discover new business goals, which can lead to a new process iteration, deepening the existing knowledge.

  The review of the process is the second step of this phase. This step aims to understand if all tasks were performed correctly and without “leaks”, such as under/overlooked features, data leakage or other failures.

  With this information, it should be possible to analyse the potential for the model deployment, improvement or need for refining the process, and take a grounded decision on the next step. The decision to deploy or reiterate the process should take into consideration the remaining budget, time, and resources for the project.

  The CRISP-DM methodology divides this phase into the following activities: Evaluation Overview; Evaluating the Results; Review Process; Determining the Next Steps.

- Deployment: Deployment of the solution.

  To proceed with the deployment, it is necessary to define how the results will be used, by whom and how often. This information allows defining a deployment plan which elaborates how the model will be used by final users, how the solution will be installed and integrated into the organization's technological ecosystem and how it will be monitored.

  Monitoring and maintenance are crucial for data mining solutions. A plan should foresee possible changes in the technological and business environment, and on how to monitor solution output performance. This takes special importance for solutions supporting decisions where changes can lead to outcomes degradation. To safeguard these situations, triggers can be set as limits for model effectiveness.

  The deployment phase can be closed with a final report and review. It can include, for instance, how the data mining process goals were met, or the new findings achieved by the process. It is also possible to enrich the phase with a project review, containing the process experience, what could be improved and other details useful for future projects.

  The last phase of the methodology process includes the following steps: Deployment Overview; Planning for Deployment; Planning Monitoring and Maintenance; Producing a Final Report; Conducting a Final Project Review.

3.2 Thesis Framework

As previous approach, before deciding which methodology would be adopted for the thesis research, considering the thesis purposes and challenges, an extensive research was performed around machine learning and data science methodologies. CRISP-DM was undoubtedly the most appropriate methodology among the methodology's literature review.

Nevertheless, this thesis proposed to go behind the "typical" machine learning research, where the goal is usually to achieve good performances and/or overcome existing metrics in models forecast, classification or other machine learning outcomes. The thesis goal was, in first-hand, to extract hidden value and knowledge from existing data which could benefit society health and well-being, and in second hand, attain the machine learning models able to support the initial goal.

Although CRISP-DM methodology was a great starting point, it was considered by our team, limited in some areas, and was felt the need to add some additional tasks and activities in a way to achieve the proposed thesis goals. The limitations were related with:

1) the model's optimization and how to both assure to achieve the best possible performance, while using resources in a smart way;

2) extract value and knowledge from data and machine learning models, and not "simply" create models which were performant or not;
3) beside doing (or not doing) a deploy of a machine learning solution, having the ability to communicate the value and knowledge extracted in the process, in a way it can benefit the business areas, whether it be business outcomes, healthcare novelties, or any other benefit.

As part of the work developed, the thesis proposes to extend the CRISP-DM methodology with three new activities, taking advantage of the recent novelties in the artificial intelligence and machine learning areas. The proposed framework can be resumed by the Figure 3.2, which highlights the three new proposed phases in an orange rectangle.

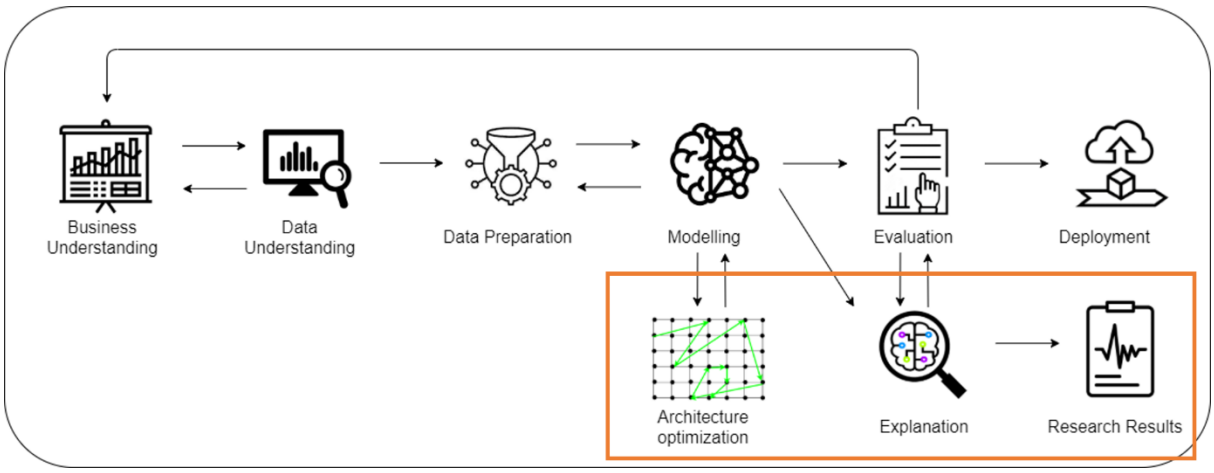


Figure 3.2 Thesis proposed framework. Three new phases (highlighted in the orange rectangle) were added to the base methodology, CRISP-DM

CRISP-DM was conceived about two decades ago by a European Union project under the ESPRIT funding with the participation of five companies, namely: Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation and OHRA. More recently, a new consortium was formed to discuss and update CRISP-DM but the results are still not published. Although there haven't been recent updates on the methodology is still considered "*de facto*" standard for developing data mining and knowledge discovery projects and have been found useful and base for countless researches. The methodology is used in a variety of areas such as engineering, education, healthcare, logistics, warfare, tourism, and much more [124] [125].

In these past two decades, and mainly on the most recent years, the evolution in the data science area has been remarkable and we can observe new algorithms, new techniques, and all kind of innovations happening with an interesting cadence. These novelties (or at least some of them) have brought opportunities that, in our opinion, where not reachable with the current CRISP-DM phases. To mention a few novelties, and the most important in our perspective: a) the machine learning architecture optimization that traditional was based in trial/error approaches has nowadays available smarter and more efficient methods; b) Explainable AI which aims to uncover the knowledge behind the machine learning models decisions. There are several studies agreeing and supporting the need, utility and benefits (some consider it an utmost need) on including these techniques on data science research and projects. In healthcare research is a subject frequently mentioned as required [126].

Taking advantage of this novelties was crucial to the goals defined in this thesis, and those we were compelled to propose an expansion to the original CRISP-DM methodology, with three new phases. The proposed new phases are presented below.

3.2.1 Architecture Optimization

The architecture optimization phase is responsible for producing the best performant model for the particular given data, considering the challenge goal(s).

Not all projects/researches which include machine learning models need to achieve the best possible performance. For some, typically not critical areas, a good performance is sufficient and sometimes having something better than simple random options can already represent an advantage. E.g.: a wrong recommendation of a movie or a product may have little impact in business. However, a better performance is always desirable as it can meet better business results.

In other cases, a very high or the best possible performance is a critical requirement, and sometimes a 99% performance, can still be insufficient. For instance, a wrong clinical decision in a surgery or in a cancer diagnose can have huge impacts in the final outcome. It is important to recall that model's performance can be expressed in different units such as accuracy, mean absolute error or AUROC (area under receiver operating characteristics curve), depending on the machine learning model characteristics and goals. These metrics are defined in the methodology first phase.

A model performance depends on a wide range of variables. It starts by the algorithm family and in the specific adopted algorithm, and from there in the particular parameters defined for the algorithm. Perhaps the simplest variation of algorithms can be found in a linear regression which has no parameters or in the popular K-means clustering algorithm, which only requires the specification of the number of clusters in which divide data. Unfortunately, most of the ML algorithms are not that simple and in a generic way, after deciding each algorithm to use for modeling, is necessary to consider a considerable number of parameters to optimize. This optimization is known as hyperparameters tunning.

A simple decision tree algorithm can easily reach more than ten hyperparameters. For instance, the decisions about the split criteria, maximum tree depth, maximum number of features to consider or max-leaf nodes, where each parameter can also take many possible values. Some of these values encompasses a continuous range, where small variations can result in big impacts, which results in combinations almost endless.

In a generic way all machine learning algorithms have hyperparameters. These hyperparameters can be divided in training and structural hyperparameters. The parameters presented above, for a decision tree, were all training hyperparameters, meaning that these options need to be defined for the modelling training phase, which directly impacts how training is performed and consequently the model performance. In the neural networks family, training hyperparameters would be, for instance, learning rate, optimizers or batch size.

The structural hyperparameters are related with the algorithm structure itself, mining the components and/or the blocks which compose the model structure. This group of hyperparameters include, for instance, the number of layers, the number of units in each layer or the activation functions which define a neural network. The concepts of hyperparameters, both structural and training, and also model parameters are best detail on the chapter 4.1 Architecture Optimization.

This proposed new phase, which we named as Architecture optimization, is responsible for addressing the challenge of reaching the best performant model, which includes choosing the algorithm family, the algorithm, the correspondent hyperparameters and parameters, while using available resources in a smart way.

To achieving the phase goal, we propose the following steps: Define the available resources for the project; Analise available strategies to architecture optimization; Select and procced with the architecture optimization; Assessing optimization results.

The methods used in this phase are discussed and detailed in the 4.1 Architecture Optimization sub-chapter.

3.2.2 Explanation

The explanation phase is responsible for allowing humans to understand the models' outputs, and how models' decisions are produced.

One of the most popular trade-offs in machine learning and data science is the binomial model performance against complexity, where usually increasing the model's complexity leads to loss of understanding/explainability. It's frequent to find complex machine learning architectures performing better that a simplest algorithm, which explains the need for adopting these solutions.

Simpler and often less accurate models such as a logistic regression or a decision tree are quite interpretable and can be described as white-boxes. In the other hand we can find the black-boxes, models which are characterised for being more complex (can reach high levels of complexity) and where it is difficult (If not impossible) to comprehend the features importance and how they related between them. In all cases explainability it's important, the difference is how to achieve it.

This proposed phase addresses how to approach explainability in a data science process. Depending on the algorithm family, the input data and the model output there are many different ways to explore explainability in models. These techniques are developed in detail in the chapter 4.2 Explanation .

Explainability can be important in many ways, such as: verifying legal and ethical constraints, bias identification, information leakage, testing model robustness, assessing societal equality and fairness, understand causality among data variables, verifying if the basis of the decisions are well fundament (informativeness), or simply to support a trust relationship between the model and the models' beneficiaries (confidence). It also important to recall that humans are generally hesitant to accept procedures that are not easily interpretable, tractable, or reliable, and at the end the humans are always the model's final users [127].

Recently, during the Covid pandemic, there were several research projects developed having as goal to support doctors during the pandemic. According to newly literature reviews most were not able to deliver value to the healthcare. The reasons for failing are varied, but many were related with data access and quality which resulted in bias outcomes. For instance, one of the research projects has used both chest radiographs taken on patients lying down in bed and standing up. Lying down scans were typically taken from patients in worse health conditions and the model learn (wrongly) to predict risk from that characteristic. In another case the risk prediction was influenced by the fonts used by different hospitals when labelling CT scans, and scans from hospitals which were dedicated to acute covid cases were classified (wrongly) as a higher risk. Both cases could have been identified with the use of model's explanations and possible misapplications and the correspondent consequences could have been avoided. Also, using explainability would eventually contribute, depending on the research direction and goals, to acquire knowledge about the disease and how it develops [128] [129] [130].

For the current new phase, we have proposed the following activities: Assess available explanation techniques to the model algorithm; Select and perform the explanation technique to adopt; Evaluate the explanation results (and decide whether a new iteration is required); Resume the findings.

3.2.3 Research Results

CRISP-DM methodology allows to define, in the business understanding phase, the goals for the process, namely the business goals and the goal for the data mining process. With a good business understanding and a clear view in what to achieve from the data mining process, it is usually easy to define these two goals when there are aligned. The limitation comes when the business goal is not simply obtaining a forecast or classification as outcome, but rather to extract knowledge and/or value from the process, as it is proposed by this PhD thesis. This is the limitation which leads to this proposed new phase.

This proposed phase is responsible to communicate the results and business conclusions obtain from the produced model, which models the knowledge, in combination with the (also proposed new) explanation phase, which uncovers the principals of the model decisions or, in other way, explains the basis for the models' outcomes.

In this new phase we proposed the followings steps: overview of the defined business goals; analyse of explanations on a business perspective; document acquired knowledge and how the results will be used, namely defining whom and how can benefit from it.

To be able to extract the best from this phase we strongly advice the tasks to be developed by a multidisciplinary team witch both combine the business expertise with the technical know-how, particularly in the ML explanation area.

Considering the nature of the knowledge discover it could be possible that, along with the obtained achievements, new business questions arise, and new processes, or even the need for collection of new data, is required for the new challenges.

This phase can be performed regardless of whether there is a deployment phase or not, as they serve complete two different purposes. The deployment delivers a technical solution while this phase delivers knowledge to be used in the business context.

3.3 Optimization Algorithms Benchmark

During the execution of this thesis, it was necessary to understand, analyse and compare different algorithms, namely possible optimization algorithms for the proposed new architecture optimization phase. This chapter presents the methodology applied for this particular challenge.

Although this was not a thesis direct goal, the literature review in the architecture optimization field revealed the lack of comparation between different strategies/approaches. This knowledge was essential for achieving the thesis proposed goals and those was necessary to develop this research to support our architecture optimization proposed new phase.

Benchmark, in a general way, can be described as the comparison of one or more products over a series of measures, which can be applied to several scientific and business areas. In the thesis specific challenge, the comparison of the architecture optimisation algorithms, the products are the implementations of the given algorithms, and the performance metrics are obtained by running the implementations on test sets [131].

It is quite common to find benchmark structured methods or processes on business companies, e.g., for comparing, assessing, and/or measuring different perspectives such as product offers or sales, personal performance, or financial ratios. Nevertheless, it becomes difficult to find processes or best practices for optimisation algorithm comparison, particularly in the AI field. After a thorough literature review, it was decided to use the recent framework presented by Vahid Beiranvand et al. in the article “Best Practices for Comparing Optimisation Algorithms”, which was created for helping researchers to design a proper comparison approach, as exactly required in the course of this thesis. To the best of our knowledge, this work is the most relevant in the area [131] [132].

The proposed framework consists of the following phases (Figure 3.3):

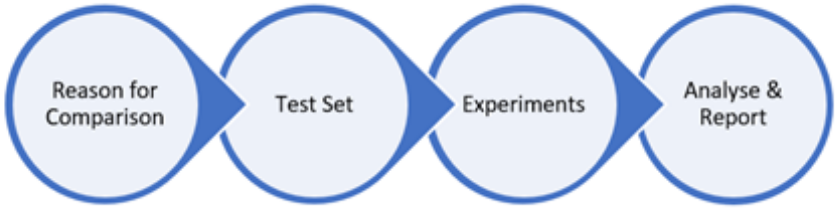


Figure 3.3 Algorithm's comparison methodology process

3.3.1 Reason for Comparison

The phase "Reason for Comparison" requires a clear definition for the reason that motivates the comparison process. The reason can be related, for instance, with an evaluation between different versions/implementations of the same algorithm, the best choice of algorithms for a specific scenario (e.g., real work problem) or comparisons between a classic algorithm and an innovative algorithm.

Besides the reason(s) for comparison, it is also important to clarify which aspects are more important for the algorithms. For instance, should the algorithm adapt/present a solution for general situations or only for specific situations (e.g.: it should include only linear problems or also consider convex universes?). What is the running time limit of each algorithm?

The reason and the relevant aspects play a critical role in the next phase definitions, which underline the need for a clear and weighted definition of this first phase.

3.3.2 Test Set

Considering the reason for the comparison, a test scenario must be defined considering that to perform a correct assessment all algorithms submitted to comparison should be applied on the same base. The test scenario may have three origins: real-world, pre-generated, or randomly generated scenarios.

Pre-generated and randomly generated datasets are often used to compare evolutions and/or new approaches to a specific reference problem, providing information about the algorithm's features and behaviours (e.g., datasets used on scientific challenges). Random approaches results can become difficult to relate to real use case problems. Real-world datasets provide specialised information on real use case problems but can present as a setback the sample size limitation or the limited access to data (e.g.: non-public datasets due to privacy reasons).

When defining the test scenario, it is important to consider the number and diversity of problems to embody, or the possibility that the dataset contains biased starting points. For instance, a small number of problems or a very specific problem can present a higher performance with a specific algorithm influencing the overall view over other algorithms, resulting in an inadequate and biased analysis. A good example of this problem is an algorithm that performs better on small numbers of features or in compact search spaces.

3.3.3 Perform the Experiments

The experiments' performance depends largely on environmental factors, namely the characteristics where experiments are conducted. The most obvious factors are the hardware and software (e.g.: CPU or GPU capacity, or the operating system) but the programmer's skills, programmatic framework, or code compilers are also non-negligible. The second relevant factor is the algorithm itself, which is exactly what the comparison seeks to assess.

To proceed with an evaluation, it is necessary to define the evaluation metrics, which will be collected during the algorithm's execution on the test set. The measures must be defined according to the first phase definitions and may be classified into three categories: efficiency, reliability, and solution quality.

Efficiency refers to the consumption of resources that the algorithm needs to run successfully. For instance, the algorithm's total running time, or the number of cycles needed to produce a solution. Reliability is related to the probability of achieving good performance on a variety of problems and can be evaluated, for instance, on the dispersion of the presented solutions for non-deterministic problems or by the success rate on deterministic situations. Finally, solution quality is evaluated considering two possible variants. In a known solution, a fixed-target goal is adopted, which evaluates the necessary time to reach the known solution; in the variant unknown solution, a fixed-cost approach is defined, executing the algorithm according to a defined cost. The cost can be, for instance, time or cycles and the evaluation will consider the error/difference between the obtain result and the best-known solution. For unknown solutions, which is quite frequent in real-world scenarios, it common to apply a fixed-cost approach, and consider for the evaluation the best-known solution or an estimative of the optimal solution (which can be done with resort to statistical techniques).

3.3.4 Analyse and Report the Results

When analysing the results, the use of graphics and tables are frequently used to assess the results obtained from the measurements and comparisons. These are powerful and versatile tools, which can improve human readability and the exploration of both generic and detailed perspectives.

Numerical tables can describe in detail the experiments measures, which offer a certain level of completeness, but it can also be overly presented making it harder to extract useful information. Summary tables are often a good solution since they are easier to read, allowing quick insights about the data. These tables can be enriched with qualitative measures that summarise numerical metrics (e.g.: run time qualified as acceptable or excessive). However, a summary comes with a cost, namely the loss of some detailed information that may be relevant.

Graphics, as visual methods, have the potential of providing better and more complete information to the reader but, for achieving the goal, it is recommended to have skills in the area or at least develop the charts considering graphical principles and best practices. A poorly drawn chart can easily confuse or mislead the reader leading to wrong conclusions. Graphics can take diverse forms, from the simplest histogram or scatter plot to a more complex box plot or convergence plot, or even more advanced charts with regions, curves, and intervals.

CHAPTER 4 - Framework Methods and Tools

The framework used in this PhD thesis research works was documented in chapter 3.2 Thesis Framework, and the correspondent phases are illustrated in the Figure 3.2. The framework is based on a well-known and proven success methodology called CRISP-DM, which this work proposed to expand in a way to achieve the research goals [118] [119].

The proposed framework is composed of nine phases, including business understanding, data understanding, data preparation, modelling, evaluation and deployment, "inherited" from the CRISP-DM methodology, which are expanded with three new proposed phases namely, architecture optimization, explanation and research results. This framework enables the implementation of data mining processes, but also the optimization of ML architectures, the exploration of model explainability in order to uncover (hidden) knowledge, and the dissemination of important insights and knowledge for the benefit of the community.

This particular chapter will not focus on the CRISP-DM phases, but rather in the proposed new phases exploring which methods and tools are available for the execution of the proposed activities. We recall that these new phases were motived not only for the specific challenges of this thesis, but also for the current CRISP-DM limitations and the opportunities presented by the recent novelties in the artificial intelligence and machine learning areas. This subject is further detailed in the section 3.2 Thesis Framework.

4.1 Architecture Optimization

Machine learning models are composed by two different types of parameters, namely hyperparameters and parameters. These two concepts are sometimes confused because they share similar designations and is not uncommon to led to misunderstandings. Those, it is important to clarify each of the concepts.

In a resume way hyperparameters are parameters which are defined by the user/programmer before the model's training. Hyperparameters are also said to be external to the model because their values won't change during the learning process. Model parameters are internal to the model, meaning these values are learned during the algorithm training process. These are composed by the internal coefficients or weights of the model found in the learning process.

One other perspective is that parameters define how input is used in a way to generate output while the hyperparameters determine the model structure and how training develops.

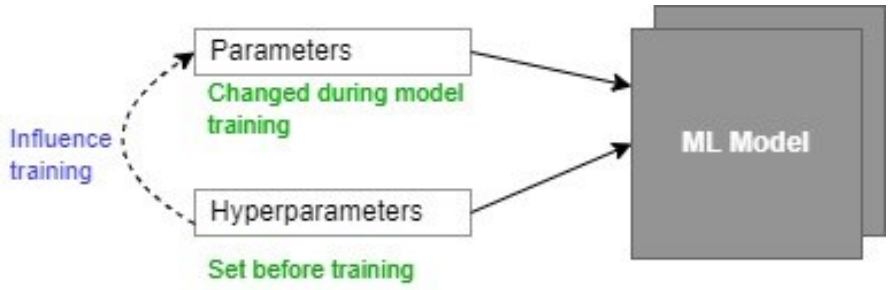


Figure 4.1 Hyperparameters Vs Parameters in a Machine Learning model

Model parameters are typically initialized to some value (it could be zero, random or based on past experiences) and then updated, during de learning process, by an optimization algorithm (e.g., gradient descent), which is an example of a hyperparameter. Other examples of parameters are the coefficients in a logistic or linear regression; weights and biases in a neural network; or the cluster centroids in a cluster algorithm. Parameters are not part of the architecture optimization phase, but instead part of the model training process.

The hyperparameters can be grouped as structural and training hyperparameters. As the name indicates the structural are related with the model's structure itself, and are responsible to define, for instance, how many layers or how many neurons will have each of the models' layers. The training hyperparameters define how training will be developed and can be determinant in the way the training universe is explored.

The hyperparameters have a known general effect in the model but is not known exactly what the behaviour for a particular dataset will be. Also, the way hyperparameters interact between them may be nonlinear which makes harder to foresee the parametrization outcome. There is currently no empirical data on which hyperparameters are crucial to tweak and which hyperparameters produce comparable performance when set to a fair default value.

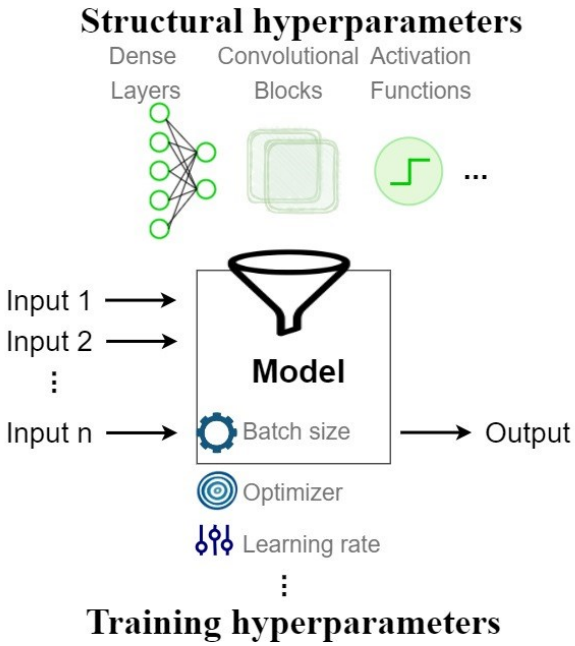


Figure 4.2 Machine learning architecture illustration

A machine learning model architecture optimization is a type of optimization problem, where the goal is to find the best combination of values for the hyperparameters.

The architecture optimization can be defined as the method of selecting the optimal set of hyperparameters to maximize model performance for a given dataset. Again, it is important to clarify that even though parameters are involved in the process, they make part of the training process which will determine the model's performance.

When performing machine learning or data science projects the most straightforward approaches to define the model's architecture consist in using an existing network structure (e.g.: a reference network) and adapt that structure or build and configure a new architecture. None of these cases guarantees the best solution for a given problem (it can be easily far from the optimal solution), but architecture optimisation is a challenging, complex, and time-consuming process. Moreover, when building and configuring a new architecture from scratch, it can be extremely difficult to achieve the optimal solution, due to the many architectural design options that require a specific level of expertise, knowledge, and effort.

One of the most dynamic areas on machine learning on the past years has been exactly how to overcome this challenge. Currently, the following main strategies are available:

4.1.1 Random Search

The random approach consists of understanding the algorithm hyperparameters, or at least focusing on the most relevant, define possible values for each of the hyperparameters and make trials (meaning training processes) on random combination of parameters. This simple task already requires a significant level of expertise. For instance, in a basic decision tree algorithm, it involves already, at least, five or six hyperparameters, such as split criteria (e.g.: Gini index, information gain), maximum tree depth (meaning the tree levels), or the minimum samples for creating a leaf node.

It gets even more demanding when it is not possible to define which of the various machine learning algorithms will be able to deliver the best performance for the particular case in study. In this case it is necessary to know all hyperparameters for all possible algorithms, as well as define their possible ranges. This demands a high level of expertise and can also lead to demanding resources consumption.

Once defined the universe of algorithms, their hyperparameters and the correspondent range of possible values, the random search consists of creating random combination of these hyperparameters and performing the model training, evaluating the achieved performance. Usually, the process is limited by a number of trials or execution time. For instance, if we take a decision tree example from a healthcare problem (e.g.: grow assessment), the search might be constrained by the number of decision trees that have been tried, or just by the allocated time to try various combinations of hypermeters.

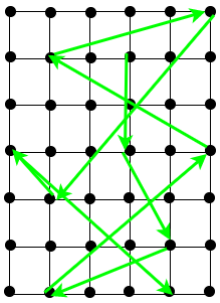


Figure 4.3 Illustration of the random search strategy

This strategy was applied in the use case documented in the CHAPTER 6 - Applied Data Science: Child's Target Height Prediction, and was one of the options explored in section 5.1.2 1D CNN NAS Testing and Measurements.

4.1.2 Greedy Search

Also known as grid search, it is an approach similar to random search which demands to define the algorithms to test, as well as the correspondent hyperparameters range. The biggest difference is that instead of using a random combination of parameters it is created a grid of hyperparameters which is then tested. The concept consists in changing only one hyperparameter at each trial and test all the possible combinations. This option also demands expertise about algorithms and algorithms hyperparameters.

This strategy usually processes the search in the universe of all possible hyperparameters combinations which can easily achieve hundreds, thousands of even more combinations, depending in the algorithm complexity. For instance, in a NN would consist in the number of layers, optimizer, activation functions, and so on. A process illustration is shown in Figure 4.4, where each dot represents a possible combination of hyperparameters.

While random search is usually faster but might miss some important spaces from the search universe, greedy is more exhaustive, but also resources eager, which may lead to an endless process.

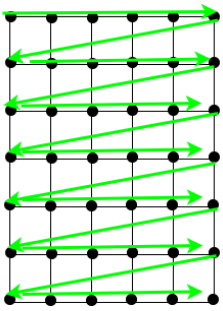


Figure 4.4 Illustration of the Greedy search strategy

Similarly to the previous strategy, it was also used both in the CHAPTER 6 - Applied Data Science: Child's Target Height Prediction and in 5.1.2 1D CNN NAS Testing and Measurements..

4.1.3 Bayesian Optimization

In a generic way, this approach uses probabilities in a way to maximize or minimize a function, which in this case consist in defining the function for the model performance. The Bayesian method, based on the Bayes theorem, considers the last combination of tried combinations to define what will be the next attempt, and this way provides a "oriented" search for the optimal solution. It basically put focus on the hyperparameters values which had provide best results, starting the new iteration in near positions in a way to find the best solution. This "smarter" search increases the probabilities of success while reducing the number of iterations when compared with the previous alternatives, and consequently reduces the resources consumption.

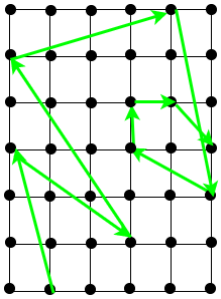


Figure 4.5 Illustration of the Bayesian search strategy

The use case presented in CHAPTER 7 - Applied Data Science: Length of Stay in Neonatal Intensive Care Units make use of this strategy to the architecture optimization phase, both for NN and regression algorithms. The strategy was also included in the work CHAPTER 1 - Neural Architecture Search for (1D) CNNs.

4.1.4 Hyperband

The hyperband concept lies on the principle of speeding up hyperparameters configuration evaluation, increasing the amount of resources spent on hyperparameter setups that promise better success while fast removing those that don't. This strategy is also known as an early stopping method.

The goal for this strategy is to be able to expand the search universe, while maintaining (or even reducing) the necessary resources. Hyperband allows to evaluate more hyperparameter combinations that other approaches which require that not all configurations run until the completion of the training process.

The process consists of defining an initial large universe of hyperparameters combinations and perform a quick evaluation on all of them using a limit amount of resources, that is, processing the training for a limited time and/or amount of computational resources. The performance of the solutions should then be evaluated, and the less performant solutions (a metric such as 60% or ¾ can be adopted) should be discarded for the next iteration. The next iteration will proceed with the best set of the solutions, allocating more resources. The process is repeated until only one configuration, the optimal configuration, remains. Figure 4.6, where the blue represents less resources and red more resources, illustrates the concept of hyperband allocating more resources to the most promising combinations.

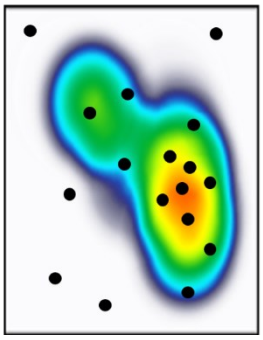


Figure 4.6 Illustration of the Hyperband search strategy

The Hyperband strategy potential was evaluated in the section 5.1.2 1D CNN NAS Testing and Measurements.

4.1.5 Genetic Algorithms

The basis for this strategy is the process of natural selection documented by Charles Darwin, which is why this strategy is also frequent refer as evolutionary algorithms.

In this concept the individuals are composed by “genes”, where the genes represent each of the options which can be taken in the model configuration. In other way a gene represents an hyperparameter for a potential solution which acquires a possible value.

The process starts with an initial population where the “genes” are initialized (a random strategy or any other strategy can be applied). All those individuals are trained, and the performances are calculated.

The following step of the process is to discard part of the individuals and generate the new generation. Usually, only half to a quarter of the individuals is kept. The new generation is based on the best previous individuals ("parents"), which will be the base for creating new individuals (child's). The child's are usually created from a fusion of a pair of selected parents where the child's genes are inherited from the parents. Figure 4.7 presents an example of crossover to generate child's, but this is just an illustration of one possible method. The new individual's generation can take other forms, such as random pick from each parent the genes to inherited or several crossovers' points. It is also possible to introduce mutations which consists of picking genes which will be mutated, meaning that they will not be based on the parents' characteristics but in other strategy (it is very common to use a random strategy). Mutations are usually adopted in a way to introduce diversity in new generations which allows to expand and bring variability to the search universe.

The new generation individual's performance is calculated, and the cycle is repeated for a defined number of iterations, where the best models will survive to the end of the process.

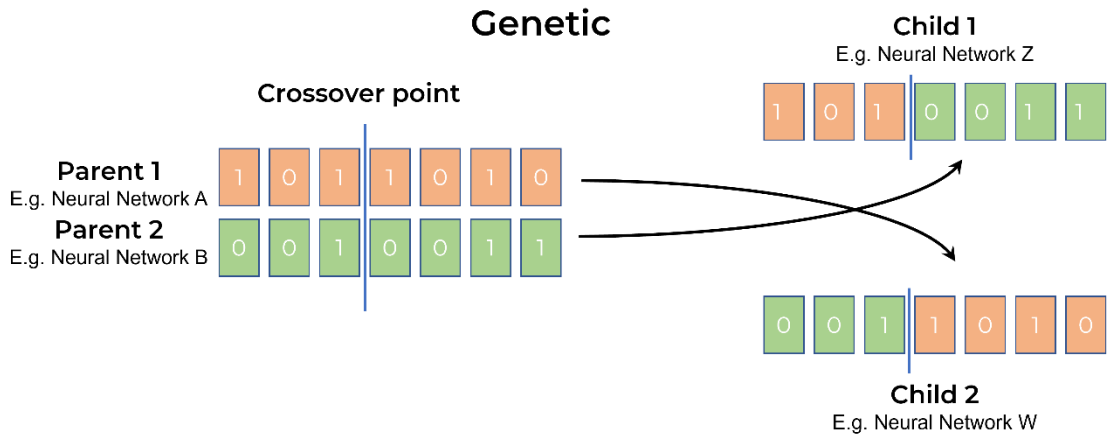


Figure 4.7 Illustration of the Genetic search strategy

One example of a genetic algorithm application is in the optimization of NNs. In this case the genes are characteristics of the NN, such as the number of layers and the number of neurons in each layer, and the parents are previously existing NNs. The descendants would be new NN, inheriting characteristics from their "parents" (Figure 4.7).

To complement the NAS algorithms benchmark performed under 1D CNN (sub-chapter 5.1 Neural Architecture Search for (1D) CNNs) was decided to also include a genetic algorithm approach. This approach was the best option for the use case presented in CHAPTER 8 - Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children.

4.2 Explanation

As previous approached in chapter 3.2 Thesis Framework, the goal for this phase is to provide explanations in how the machine learning model produces the outputs and what are the bases for the model's decisions. In other words, to understand the features importance and how they related to each other in a way to produce the model output. Figure 4.8 illustrates the concept, presenting in the left a ML model which if "fed" with inputs and produces an output as a black-box, meaning that the bases for the output production are not presented. The right plot presents the advantage of introducing the explanations, which uncovers the basis for the model's decisions [133].

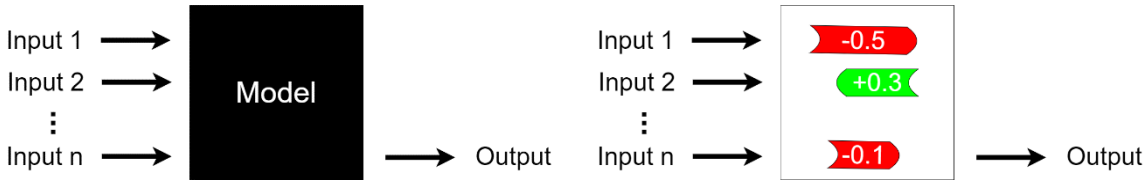


Figure 4.8 Illustration of the Explanation concept

The explainability of a model can be achieved in different ways and highly depends on the model algorithm. For instance, in a linear regression it is possible to interpret the impact of features in the model outcome just by looking at the generated linear regression expression.

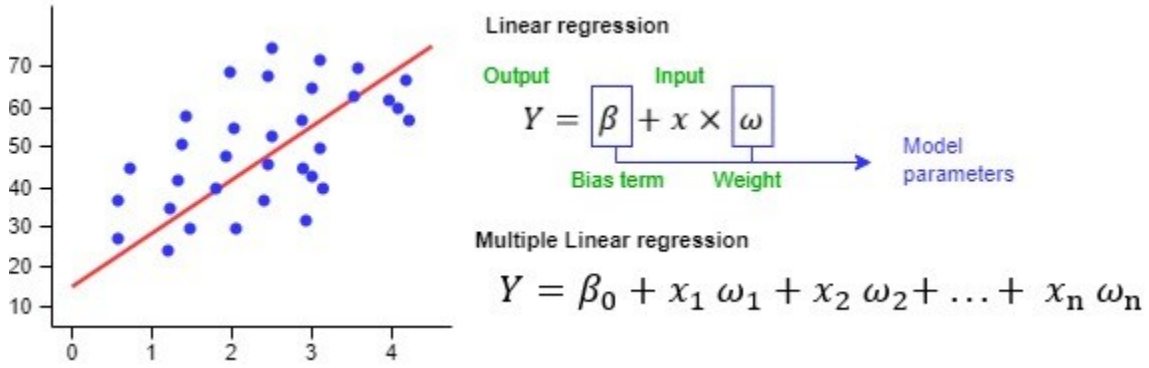


Figure 4.9 Linear regression illustrative plot and expressions

In the mathematic “linear regression” expressions presented in Figure 4.9 “Y” represents the model output, “β” is called the bias term or interception and it is a constant term learned in the training process, “x” represents the model features observations (feature 1,2…N) and ”ω“ the weight or coefficient associated with the correspondent feature, also defined by the training process. The ”ω“ describes the weight of “x” in the final output, where a higher weight represents a higher relevance/influence. The importance of a feature can be seen as the absolute value of the t-statistic value.

For the sake of simplicity, and as an example, in a binary classification task each weight can see as a percentage of contribution to each of the classes (0 or 1).

The main drawback for this algorithm is that it can only model linear relationships, and in real world problems that is not common which usually results in a model with not so good performances.

In a model based on a simple decision tree it is also possible to understand the features importance and interactions just by looking on how the tree was built. A decision tree algorithm consists of a cycle where each time the model chooses the feature which best splits the dataset. There are many ways of evaluating that decision (e.g.: cross-entropy or Gini index), and to define the stop split criteria (e.g.: max tree dept or inability to lower Gini index), but that topic will not be further discussed as it is not relevant for this chapter. It is this process of generating the tree that makes it easy to interpret a decision tree diagram and to directly extract both features importance as well as how the model output was produced. It is important to highlight that when using variables with different ranges such as continuous features both with a wide and short range, discrete or binary options, it is necessary to assure fairness in the analysis, as feature with a higher cardinality tend to be benefited on their relevance when analysing a three decision algorithm.

The Figure 4.10 illustrates a basic decision tree classification example where it’s possible to obtain the explainability of the model.

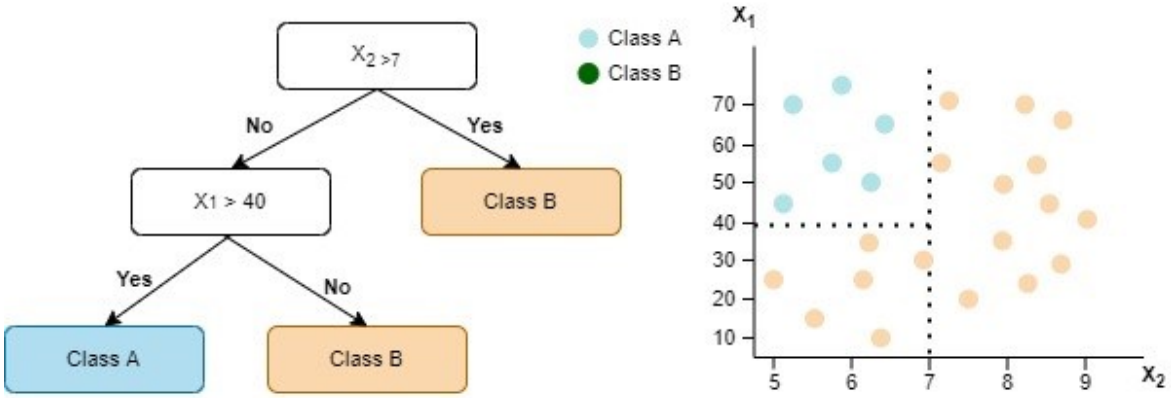


Figure 4.10 Decision tree algorithm illustrative explanation

There are other models which present explainability just by the algorithm itself (the so-called interpretable models by design), such as K-Nearest Neighbours, Bayesian models, rule-based learning model, general additive models, or Lasso and Ridge regressions.

With the algorithms complexity grow it's necessary to adopt other strategies for explainability. For instance, a random forest algorithm frequently achieves better performances that a basic decision tree but would be harder to analyse. Although the algorithm is still composed by decision trees, the combination of many decisions tree (it can be dozens, hundred, or even more) would require looking at each decision tree and extract associations between them. For that type of challenges, it is required another means for explainability using additional techniques.

These techniques are usually known as interpretable models and model interpretability techniques, or transparent models and post-hoc/post-modeling explainability. These techniques are approach in the following sub-chapters.

4.2.1 Leave One Co-variate Out (LOOC)

Also known as drop column importance, as the names suggests, resorts to remove iteratively features from the model (one at each time) and analyse the impact in the model's performance.

It is an interesting technique, model agnostic, which is able to provide a quick explanation about the predictive power of each feature, and for instance overcomes the decision tree drawback related with features with a higher cardinality. The disadvantage is the resources consumption as it is necessary to train a model multiple times, and the effect caused by collinear features which can lead to wrong conclusions. In a collinearity case, the model training tends to "substitute" the excluded feature by the information included in the features which are collinear related [134].

4.2.2 Permutation Importance

This technique, illustrated in Figure 4.11, allows to extract feature importance in a quite fast and simple way. The technique consists of "experiments" made with the validation data, where the values of one of the model's features are randomly shift while maintaining the remaining features and target label. The process is performed to each of the features (separately) assessing how the change affects the model performance. The degree of impact in the model performance defines the feature importance [135].

The process is very intuitive, model agnostic, and the ML model doesn't not to be retrained. As the LOOC it is also affected by the effect of features correlation. Plus, the randomness of the process can lead to randomness in the outcomes, or even result in combinations of inputs which are impossible. Taking for instance, an input feature presenting the age of a father and a son: the randomness of the process can result in a record where the son is older than the father, which off course distorts the analysis.



Figure 4.11 Permutation Importance method illustration

4.2.3 Individual Conditional Expectation (ICE)

The Individual Conditional Expectation focus on the model outcome impact produced by an individual feature value change, for instance, with an increment or reduction of 1 unit. These results are analysed considering several prediction instances produced by the trained model, using different values for the feature. This method is limited to regression problems and can only analyse one feature at a time, which limits the application of the method. On the other hand, is a very intuitive process to understand.

In the Figure 4.12 we find the left plot presenting a feature (X_1) which reflects a linear relation with the model output (Y), while the right plot exemplifies a non-linear relation between the feature (X_2) and the model prediction.

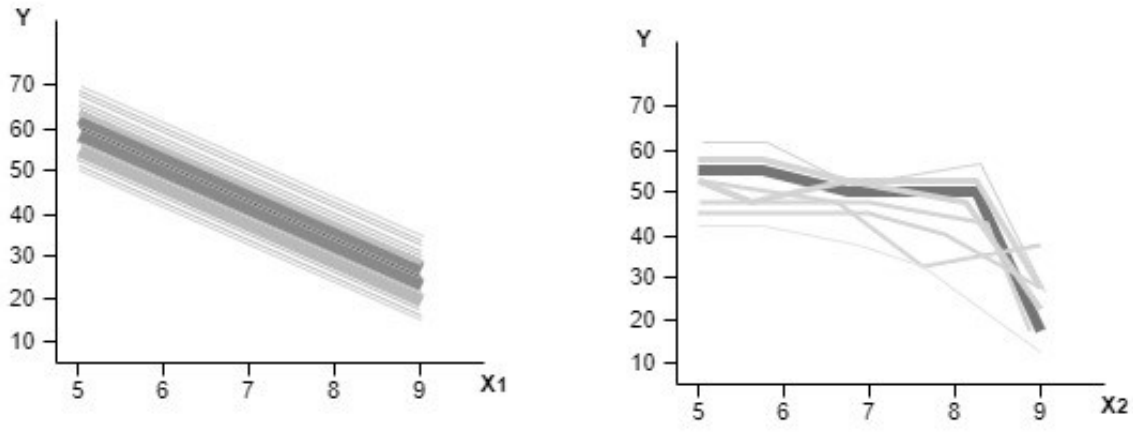


Figure 4.12 Individual Conditional Expectation illustration, with a linear and non-linear example (respectively X_1 on the left and X_2 on the right)

There are some variants on the ICE to make easier to capture information such as centred and derivative ICE plots. The base concept it's always the same, but they diverge in how the data is presented.

4.2.4 Partial Dependence Plots (PDP)

This method, sometimes just called by PDP, is similar to the Individual Conditional Expectation, allowing to understand how features value changes affect model output (global explanation), but expand the information a little further. On a PDP, a single feature is the average of the lines of an ICE plot. Again, the trained model is used to forecast several instances with repeated variations in a single or a pair of features (the method could reach until three features, what is already hard to understand). The method outputs a confidence interval for the predictions (a difference when compared with the previous method).

Figure 4.13 presents two plots. The plot in the left highlights the relation between a feature (X_2) and the model outcome (Y) with the confidence intervals as a good example of the application of a partial dependence plot to one feature. The plot in the right shows an example of a 2D partial dependence plot, and how the two variables related with the model forecast.

In a resume way ICE presents the functional relationship for individual observations while partial dependence plots show an average effect. Both are quite easy and intuitive methods and allow a good understanding of the features relationships while not resources intensive (no need for model's additional training). The biggest drawbacks for the technique are the limitation of presenting (at maximum) three dimension representations (not a method limitation but rather a human inability to understand more than three dimensions), and the fact that it assumes features independence, as LOOC and permutation importance methods. When dealing with correlated features, Accumulated Local Effects (ALE) plots, which are based on the same idea but use a different calculation formula, can help to overcome this drawback.

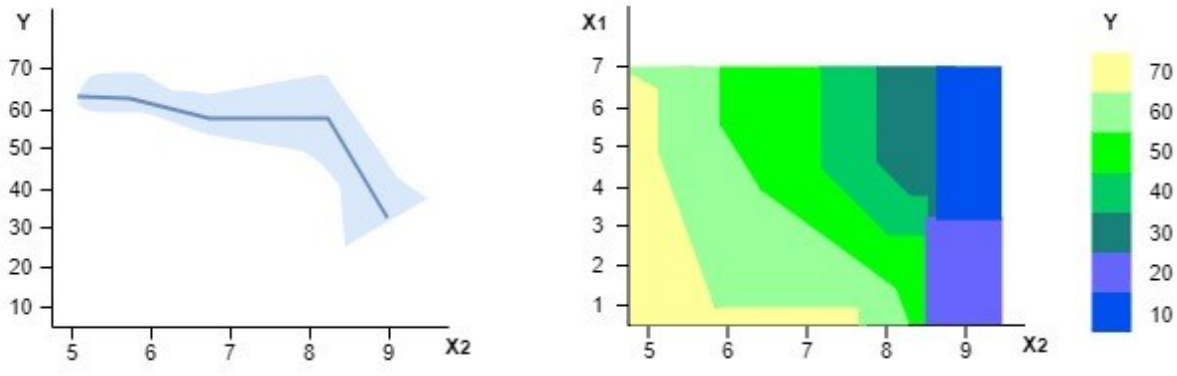


Figure 4.13 Examples of Partial dependence plots using 1D (left) and 2D (right), respectively for analysing one or two features

4.2.5 Surrogate Models

Surrogate models are also known as approximation models, or global surrogate and, as the name suggests, the concept behind involves replacing the original model (the trained model) by another model. This technique consists in training an interpretable model (for instance a decision tree) using the same data used for training the original model, but having as output targets the predictions produced by the complex model (instead of the real data labels). The goal is to produce a model with a high degree of explained variance (measure for instance with the R-squared measure), meaning the surrogate model approximates the original trained model. The technique presents as its greatest advantage the possibility of using any algorithm for training the original model and any interpretable model to explain the first (model-agnostic method) which offer a great flexibility. Also, it allows to extract both local and global explanations. The main drawback is the interpretable model's potential inability to produce performances that are consistent with the original model [134].

4.2.6 Local Interpretable Model-Agnostic Explanations (LIME)

LIME delivers a concept similar to the surrogate models, but instead of focusing in the entire model if focus on single features, training explainable models for each of them. This process it's also known as local interpretability or local surrogate models as the explanation is based on "local neighbourhoods" .

The concept was first introduced by Marco Tulio Ribeiro, et al. in 2016 in the ACM's Conference on Knowledge Discovery and Data Mining [82].

The process consists in selecting the instance to explain and create new samples making small variations to the instance input data, a process also known as perturb the input around its neighbourhoods. For instance, considering a natural language processing problem the base expression (instance to explain) could be "I love that picture" and the neighbourhood / variations could be expression such as "love that picture", "I love that", "I picture", or "I that picture".

That new instances are then submitted to the original model and outputs are saved. A new dataset is created with the samples and the correspondent inferred labels, considering how similar they are with the base instance in analysis. A kernel function or a distance function should be used to evaluate similarity. That new dataset is then trained with an interpretable model (such as decision trees or Lasso) and the results are compared with the instance to explain.

Similar to surrogate models, the benefit of this approach is the flexibility to support any original model regardless the used algorithm with any interpretable model (that is why the name of the technique states it is "model-agnostic"). Also, it is a method able to process tabular data, text or images. Presents as drawback the fact of presenting only local fidelity, as the technique does not provide model global explainability, and the fact that it can be challenging to define the neighbourhood because the term "neighbourhood" can be ambiguous, particularly for raw data. The explanations can be instable due to the sampling process which can also affect the trust in the explanations.

4.2.7 SHapley Additive exPlanations (SHAP)

SHAP is a method based on Shapley values which introduces a new approach to explainability, allowing to breakdown a single prediction. It also presents other associated resumed and aggregated information which allows to explore the produced model in a global way.

In a resumed way, SHapley Additive explanations, also known simply as SHAP, were introduced in 2016 by Lundberg and Lee and later expanded by Chen, Lundberg and Lee. This technique is based on a game theoretic approach and is used to explain predictions, namely by computing the contribution of each feature to the prediction [136] [137].

The method calculates the model "overall" base value for the model's prediction (the average of all predictions) and for each single prediction is able to decomposes the output, explaining how each feature has contribute for that particular prediction. The features, meaning the feature value for that particular instance in evaluation, can contribute for the increment or reduction of the prediction when compared with the base value. Usually, the features which increase the prediction value are presented in the red colour, while the features contributing for the reduction are presented in the blue colour. The length of the feature bars indicates the dimension in which the feature affects the output.

Figure 4.14 shows an example where the model outputs for the instance in evaluation predict the value of 50,41. The feature X_1 and X_2, or the be more exact the value for the features in the particular instance which in this case are respectively 10,12 and 0,88, have contributed to increase the models output and the features X_7, X_3, X_5 and X_{10} have contributed to reduce the forecasted value. The features are presented according to their contribution, where higher contributions are presented nearest the model output, meaning X_2 has a higher contribution that X_1, and X_7 has contribute to reduce the output more than any other feature. In a healthcare context, the variables X might be for instance weight, gender, a specific medication or treatment, and the prediction/outcome might be the number of hospitalization days.



Figure 4.14 SHAP prediction instance example

Beside this explanation for a particular instance based on a regression illustration problem, SHAP offers many other resources, such as, explanation for feature effects allowing to understand how different values for the same feature affect the model, how they related with other features, the features importance, the dimension in which features affects the output, words contribution in natural language processing models, extraction of the most influential areas in images (image features) or in computer vision processing challenges, among others tools.

This technique covers many machine learning algorithms (including deep learning which is often approached as DeepSHAP), presents a diversity of tools and provides both local and global interpretability, which makes it a very interesting technique in the AI explainability area. The disadvantage is the fact that it can be resource intensive (high computation load).

This technique was applied in the use cases “Applied Data Science: Child’s Target Height Prediction” and “Applied Data Science: Length of Stay in Neonatal Intensive Care Units”. Sections 6.9 Explanation and 7.9 Explanation delve deeper into the implementation's specifics.

4.2.8 Learned Features

In a generic way the neural networks can learn abstract features, concepts and relationships from the inputs provided to the network. Particularly the deep neural networks have the ability to extract/capture features using the hidden layers which reduces the need for feature engineering, and the need for hand-craft features, which are often complex, time-consuming, “imperfect” (meaning some knowledge is not captured) and require expertise knowledge.

For instance, in a convolutional neural network the first layers learn simple features such as edges and simple textures, while the latest learn features such as objects or parts of objects. Finally, the fully connected layers use these features to generate the outputs (classification or regression problems).

This technique consists in understanding the features learned by the NN. It is possible to understand the learned features using the concept of feature visualization which consists in evaluating the activations of the units which compose the NN (e.g.: neurons, features or activations maps) for a defined input. For instance, in a CNN it’s possible to inspect each of the layers and features maps and understand the identified forms (edges, figures, …). Some of those forms can be easily recognized and other can be complete abstract for a human [138].

One other method to understand learned features is by network dissection. The feature visualization allows to observe features which were extract by the NN, but that doesn’t mean that the forms or relations are used for generating the output. Network dissection consists in submitting new human-labelled inputs to the trained network and evaluate the NN activations for the label’s concepts, to understand which concepts were captured and have contributed to the NN output. For instance, in a CNN which classifies objects and considering a bicycle would be possible to verify which forms contributed to the object classification as a bicycle. This way would be possible to understand if the classification was made based in forms such as a bicycle wheel, a cycle track or the rider's equipment. This kind of information is crucial for both understand and audit the model’s decisions [139].

These methods allow to understand how NN work and were, for instance, responsible for allowing the scientific community to understand how features and concepts are captured in each of the NN layers. In the other hand, many visualizations of features do not correspond to human interpretable concepts which makes it difficult to understand. Plus, a network dissection requires huge, labelled datasets with concepts which were captured by the networks, and which leads to extensive efforts and resources. It's important to keep in mind that NN are composed of numerous units, which easily expand to hundreds, thousands, or more, meaning that the regions to observe grow exponentially.

4.2.9 Saliency Maps, Grad-Cam and Variations

The saliency maps are typically used for analysing image and for presenting the contribution of image pixels for the NN output. This technique can be also called as sensitivity map or pixel attribution gradient-based attribution methods.

The simplest way to apply this technique is based on excluding/occulting parts of the input (for instance break a picture or a text into pieces) or produce variations/perturbations in the input and understand which parts of the input most contribute to the output. It's basically the concept of LIME, explained in chapter 4.2.6 Local Interpretable Model-Agnostic Explanations (LIME), but applied on NN.

Another method for examining areas of interest involves the use of a gradient-based method which consists of computing the NN gradients generated by the network with respect to the input features, when producing the output. There are many gradient-based methods which basically differ in how the gradient is computed. Probably the most popular method is the Grad-CAM, which we explain in detail as a reference to how these methods are processed.

Gradient-Weighted Class Activation Map, generally abbreviated to Grad-CAM, is commonly used in image processing (for instance in CNN layers) and (usually) uses the last convolutional layer to classify the pixels of the image relevance to the CNN output. Bear in mind that this concept is usually applied in the last convolutional layer because the last layers are the ones able to learn features objects or parts of objects. Still, the technique can be applied in any of the NN layers.

The Grad-CAM generates a heatmap which encapsule the pixels relevance to the output. Overlaying this heatmap with the original image it is possible to sign the regions of interest which allow to understand the principals of the NN output. The Figure 4.15 summarises graphically the Grad-CAM process [140].

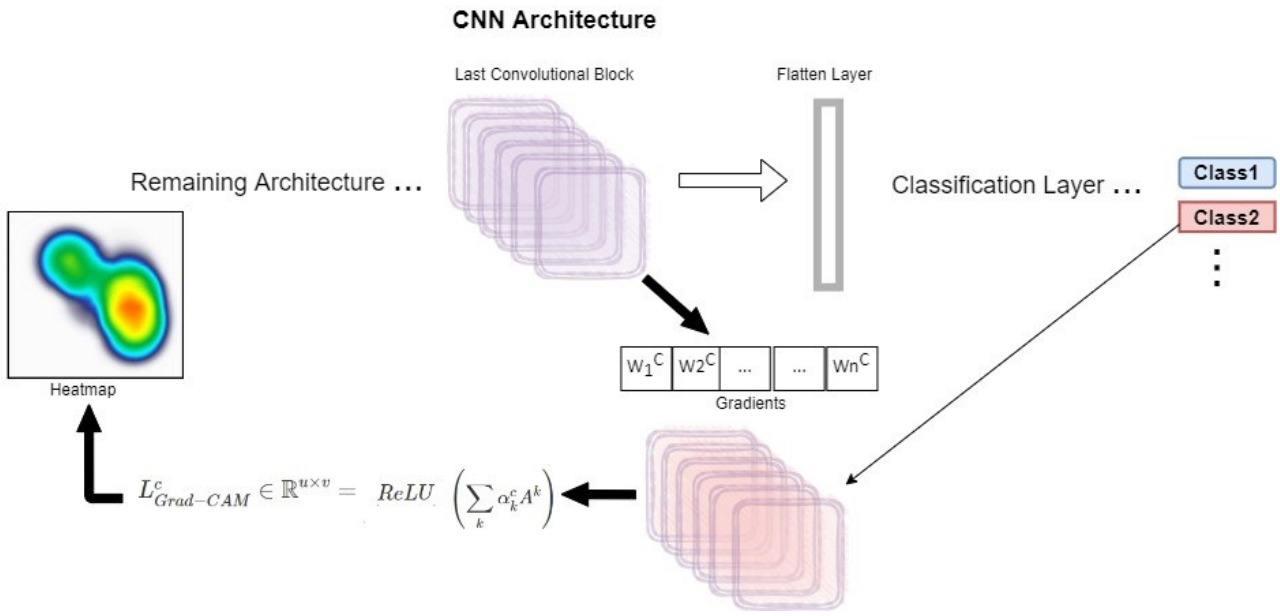


Figure 4.15 Grad-CAM method illustration

As previous approach there are many variations on the gradient-based method. Some of the most popular are Class Activation Mapping (CAM), Guided Back propagation, Grad-CAM++, Guided Grad-CAM, SmoothGrad, Score-CAM or DeepLIFT (Deep Learning Important FeaTures). They all share the same principles but differ in the way gradients are computed.

The Grad-Cam method was applied in the use case "CHAPTER 8 - Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children", but with some innovations presented by the thesis. The technique implementation is described in section 8.9 Explanation.

4.2.10 Resume

The previous sub-chapters covered the main AI explainability strategies. This section offers a summary of the characteristics of the various techniques to assist researchers in choosing which technique to use for the specific challenges.

There are other explanation methods, in fact there are several more if we include all variations for each family method, but in this chapter, was decided to present only the most promising and most used methods for explainability in data science and machine learning. The AI explainability is still an emerging area and it expected to continue to have an expressive develop in the coming years, with the likely emergence of new techniques and methods.

Table 4-1 AI explainability techniques resume comparation

Explainable Technique	Model Agnostic	Explanations	Features Coverage	Model Retraining	Drawbacks
LOOC	✓	Global	Tabular	Multi-times	One feature at a time Assumes features independence
Permutation Importance	✓	Global	Tabular	No	One feature at a time Randomness outcomes Non-realistic combination of inputs
ICE	✓	Local	Tabular	No	One feature at a time Limited to regression
PDP	✓	Global	Tabular	No	1 to 3 features at a time Assumes features independence
Surrogate Models	✓	Global and potentially local	Tabular	New interpretable model	Difficulty in produce an approximate surrogate model
LIME	✓	Local	All types	New interpretable model	Explainability outcome's instability
SHAP	✓	Local and global	All types	No	Resource consumption
Learned Features	✓	Local	All types	No	Human interpretable sometimes difficult Requires big, labelled datasets
Saliency Maps, Grad-Cam and Variations	✓	Local and potentially global	Mainly image	No	Complex implementations and requires expertise knowledge

4.3 Research Results

The research findings from a data science research initiative can result from a goal or research question defined on the business understanding phase, or simply because the data science process has disclosed relevant knowledge which the team has decided as relevant to document and/or share.

Starting with a hypothetical situation as an example, may be the most effective way to grasp the relevance and how to develop the research results phase. Let's consider a data science project which analyses electrocardiogram waves and where the challenge is to classify that ECG signals in class A or B. Following the proposed framework, a performant model would be developed, and the explanation phase would be able to analyse the foundations for the model's conclusion/outcomes, specifically in terms of what would be the differences between waves that would result in the classification into class A or B. This information would support the model validation as well as the model outcomes and help to deliver a useful technical solution. However, aside from the introduction of the new system, no new business knowledge or value was created in the electrocardiogram field! Although it is possible to identify the difference between waves, the business meaning hasn't been extracted, or to be more exactly the differences haven't been evaluated in a clinical perspective.

The research results phase is responsible to join together the technical knowledge with the business expertise and. In the given example, would be able to map the differences in the ECG signal to the exactly functions which the heart is performing and understand in practice which would be the alterations in the heartbeats as well as the potential cardiac consequences. To achieve this generation of knowledge it is essential that the team responsible for the task is multidisciplinary and incorporates both business and technologic knowledge.

Beside the team skills it is also suggested to document the acquired knowledge following some guidelines, namely:

- Define the audience and their background;
- Overview the research aims and/or the research question as well as the relevance of the study and findings;
- Describe the population and the data considered for the study. It is also important to document any possible limitations imposed by the population and/or data (e.g.: a particular age range, single ethnicity individuals, existing bias);
- Document the knowledge according to the target audience and their characteristics;
- Explain what is the add knowledge to the business area;
- Clarify how the generated knowledge can contribute to improve the business area;
- And finally, document eventual new questions that may have arisen from the process.

CHAPTER 5 - **Derivative Works (secondary outcomes)**

This section is dedicated to research work which was developed but did not made part of the initial PhD thesis scope. However, the need for tools and/or developments to achieve the proposed goals, and/or the lack of existing solutions forced the developed of new solutions. All the works presented in this chapter were crucial to support the development and research presented in the following chapters, which are focused on the thesis main goals.

Although these activities are not a “primary” output from the thesis, in our opinion, the relevance and value of the created solutions make it a “secondary” product of this thesis. Following this principle, it was decided to document these works in the current chapter.

The created solutions where shared with the community in order to benefit future research or even to be used in professional or commercial settings.

5.1 Neural Architecture Search for (1D) CNNs

In the field of sensors and signal acquisition, in areas such as industrial, clinical, or environment, it is common to find one dimensional (1D) formatted data (e.g.: electrocardiogram, temperature, power consumption). A very promising technique for modelling this information is the use of One Dimensional Convolutional Neural Networks (1D CNN), which introduces a new challenge, namely on how to define the best (optimized) architecture for a 1D CNN. This chapters addresses the concept of One Dimensional Neural Architecture Search (1D NAS), an approach that automates the search for the best combination of Neuronal Networks hyperparameters (model architecture), including both structural and training hyperparameters, for optimising 1D CNNs. Please find further details about hyperparameters in sub-chapter 4.1 Architecture Optimization .

The work includes the implementation of search processes for 1D CNN architectures based on five strategies: greedy, random, Bayesian, hyperband and genetic approaches. Data collections and data analyses are considered for each strategy scenario. For the analysis of the distinct strategies, 125 experiments were conducted, followed by a thorough evaluation from multiple perspectives, including the best-performing model in terms of accuracy, consistency, variability, total running time, and computational resources consumption.

The research documented in CHAPTER 8 - Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children, or, more specifically, the challenges encountered in the optimization of the required 1D CNN which are detailed in sub-chapter 8.7 Architecture Optimization, led to the need for this work.

Also noteworthy is the fact that the work described here was previously published in a Quartile Q1 journal under the title Neural Architecture Search for 1D CNNs—Different Approaches Tests and Measurements. With already a few citations, the community has acknowledged, in our perspective, the importance of this work [3].

5.1.1 Introduction

In the past years, the field of smart sensors, Internet of Things (IOT), artificial intelligence, and the combination of these fields has made enormous progress. Specifically in the machine learning area, Artificial Neural Networks, Deep Neural Networks (the most common techniques in the deep learning field), and Convolutional Neural Networks are concepts in vogue, reflecting a growing scientific interest and the consequent contribution from the scientific community [141] [142] [143].

CNN is a specific class of deep neural networks and perhaps the most popular algorithm among the deep learning area. To understand this popularity, there is a need to understand the deep learning and CNN's main concepts.

Before the emergence of deep learning based algorithms, the traditional machine learning techniques were limited to process data in raw format, which required careful feature engineering and extraction expertise to transform natural data into tabular forms (handcrafted features space). This process can achieve levels of complexity that, beside the necessity of high expertise knowledge, often lead to information loss. Deep Learning methods allow models to discover representations from natural data, such as pixel values in an image, replacing the handcraft and "imperfect" human data transformation [144] [145].

The Deep Learning concept was first introduced in 2006 by a group of researchers of the Canadian Institute for Advanced Research (CIFAR). It refers to models composed by multiple processing layers, where each layer can learn representations with multiple levels of abstraction, capturing the full complexity within the data. The existence of several layers in the network leads to the origin of the "deep" term when referring to this type of models [145] [146] [147].

Deep Learning has recently been used successfully in many fields such as visual object recognition, speech and audio processing, natural language processing, object detection, among others [147] [148] [149] [150].

Convolutional Neural Networks

A Convolutional Neural Network is a deep feed-forward network, that extends a "classic" artificial neural network (e.g.: perceptron or feed forward NN) by adding more layers (deep learning), including the convolutions blocks. The term "convolutional" comes from the implementation of convolution blocks in the network. The first appearance of this type of neural network dates to 1980 when Kunihiko Fukushima introduced the concept of convolution and down sampling layers, and to 1990, when LeCun et al. published a paper presenting the principles of the CNN modern framework [151] [152] [153].

The typical structure of a CNN is composed of three layers: convolutional, pooling, and fully connected layers. These layers make use of several components/techniques, such as convolutions, activation functions, pooling, dropout, batch normalisation, fully connected blocks, among others, that can be combined in multiple ways [151] [153] [154].

On the CNNs architectures, the convolutional layers are responsible to process the feature extraction tasks from the input, while the fully connect blocks process the output in the form of classification or regression. Figure 5.1 presents an illustration of a CNN architecture [155].

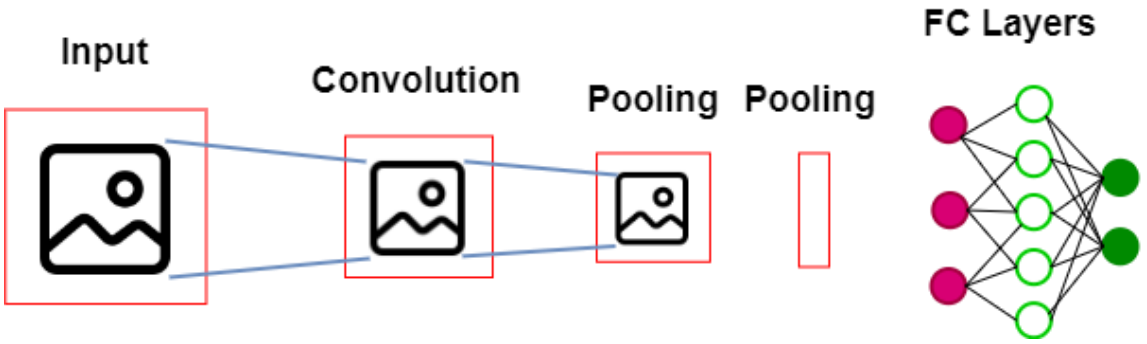


Figure 5.1 Illustration of a CNN architecture

CNN Applications

The use of CNNs has dramatically improved the state of the art on image processing (including image classification, image semantic segmentation and object detection) and, nowadays, is still the main reason for their popularity, having many successful contributions among diverse areas such as agriculture, healthcare, transportation and weather forecast. Some good examples of CNNs applications are presented by A. Kamilaris and F. Prenafeta Boldú on farming disease detection and fruit counting, or by Xiaolei Ma et al. which applies CNNs to traffic prediction in the transportation area. Some other greats examples, in the healthcare area, were introduced in the sub-chapter 2.1 Data Science in Healthcare [151] [156] [157] [158].

Besides the application of CNNs on image scenarios (data in 2D format), with several popular architectures such as AlexNet, VGG, GoogleNet, ResNet and Inception, and worldwide image processing competitions (e.g.: ILSVRC - ImageNet Large Scale Visual Recognition Challenge and VOC – PASCAL Visual Object Classes), CNNs structures make it possible to process other types of data, such as natural language, video, speech or audio [159] [160] [161] [162] [163].

Though the flexibility of the CNNs, given by 1D, 2D or 3D convolutions, and the widespread use of 2D convolutions, it's not common to find 3D convolution models, which probably can be explained by less availability of 3D data. 1D convolution models, despite the availability of 1D data, also have a significant lower popularity [164] [165] [166].

There are even situations where 1D data has been converted/reshaped to 2D format, so 2D CNNs can be applied. On the recent work about air quality presented by Mirche Arsov et al., with the title "Multi-Horizon Air Pollution Forecasting with Deep Neural Networks", the authors reshaped 1D data to 2D in order to apply a 2D CNN to their work [167] [168].

Despite the scarcity of 1D CNN research, some recent research works can be found, e.g. the work developed by R. A. Osman et al., on interference avoidance on device communications using 5G communication; the advanced diagnosis of soft faults on motor power cables by H. Kim et al.; the research "Estimation of Tool Wear and Surface Roughness Development Using Deep Learning and Sensors Fusion" presented by P.M. Huang and C.H. Lee; the monitoring of the spatter behaviours on acoustic signals developed by S. Luo et al. Examples of 1D CNN in healthcare as even more rare.

Although all the research's cited before applied 1D CNNs on their challenges, neither of these recent works used Neural Architecture Search (NAS) techniques (some don't even discuss the optimisation subject), resorting instead to the "traditional" grid search or trial-and-error approaches. This is the research gap that led to the need for this research [154] [169] [170] [171] [172].

There are several advantages in using 1D CNNs, such as good performances on a limited amount of data, low computational complexity (when compared with 2D CNNs and other deep learning architectures), faster training process, and a good ability to extract relevant features in sequence data and time-series (e.g.: signal data). Serkan Kiranyaz et al. in the article "1D convolutional neural networks and applications: A survey" discuss in detail these advantages. The research works previously presented as state of the art are also good examples on how to benefit from the implementation of 1D CNNs [166] [169].

Nevertheless, the number of published articles on 1D and 3D convolutions is pretty narrow when compared with 2D and, to the best of our knowledge, there are no reference architectures for any of the cases. There are great opportunities to explore 1D and 3D CNN in areas such as: medical 3D images (like computed tomography, magnetic resonance imaging or ultrasound) and object identification or segmentation; 1D signals, such as electrocardiograms (ECG), electroencephalograms (EEG) and electromyograms (EMG). On these areas, there is a lot of data available, which allows further research considering different CNN applications [164] [167].

The Need for Optimised CNNs

The success of a CNN, and in general for any neural network, depends on finding an architecture to fit a given problem. The most straightforward approaches are to use an existing network structure (e.g.: a reference network) and adapt that structure, or build and configure a new architecture. Any of these cases does not guarantee the best solution for a given problem (it can be easily far from the optimal solution). But architecture optimisation is a challenging, complex, and time-consuming process. Furthermore, when building and configuring a new architecture from scratch, it can be extremely difficult to achieve the optimal solution, due to the many architectural design options that require a specific level of expertise, knowledge, and effort [173].

A CNN architecture is defined by a set of hyperparameters, namely structural and training hyperparameters. The structure hyperparameters are the layers, number of units in each layer, filter and kernel definitions, among others. The training hyperparameters are the ones that define the training process, such as learning rate, batch size, initial weights or number of epochs. Figure 5.2 presents an illustration of the variables included in the CNN definition [174] [173].

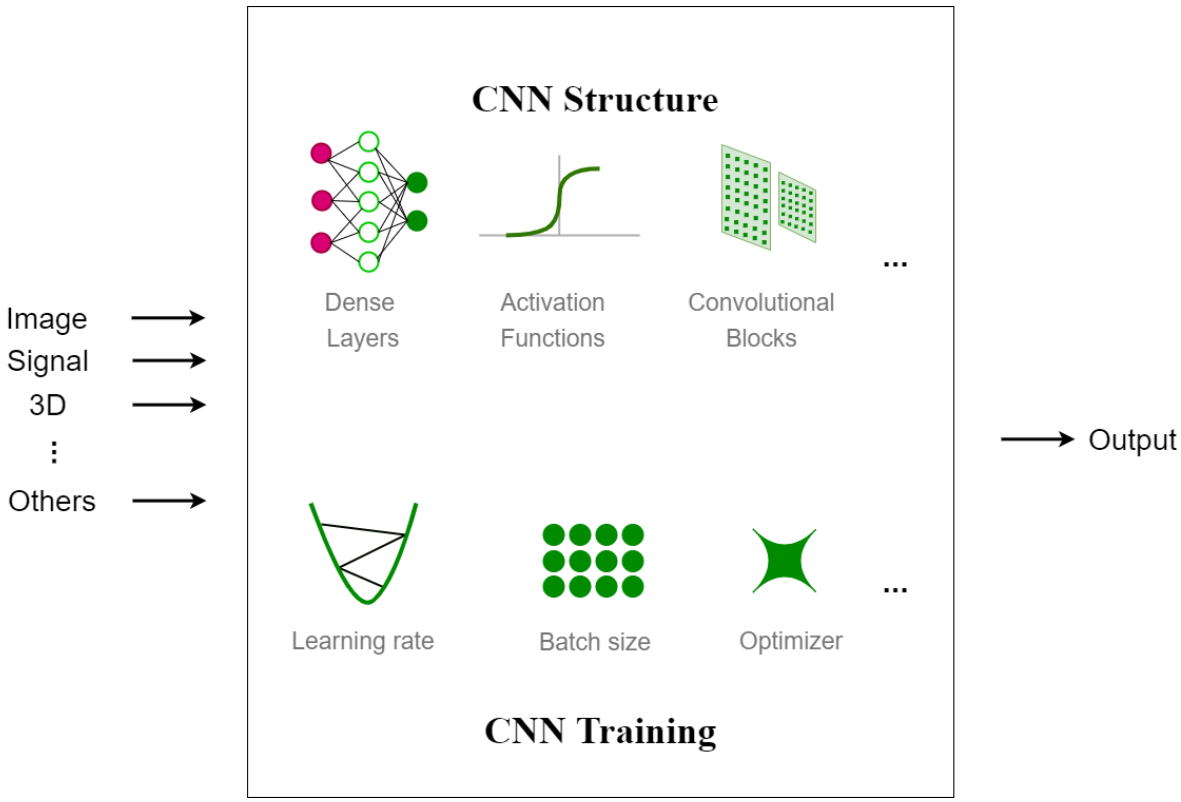


Figure 5.2 An example of a CNN architecture definition (including hyperparameters)

The performance of a CNN is critically sensitive to the architectural design and differs for each problem/challenge, leading to a difficult task of finding the appropriate hyperparameter values and their combination. It becomes even harder to understand network hyperparameters relations and their impact on the model's performance.

Since there is no mathematical formula for calculating the appropriate hyperparameters combination, it is common to find the grid search or random search as strategies to optimise the architecture. These strategies present some drawbacks, such as high usage of computational resources and non-directional ("dummy") search for the optimal solution. Even with these strategies, there is also no guarantee that the resulting model is the best solution. More recently, Bayesian optimisation methods and genetic algorithms have been considered by the scientific community as approaches to optimise the architecture, but still with quite limited experiments [173] [175].

Neural Architecture Search for Neural Network Optimization

Figure 5.3 illustrates the NAS concept that describes the process of automating the search for the optimised design for neural networks through a search strategy. This concept was introduced in 2017 by Research at Google Zoph & Le 2017 [176] [177].

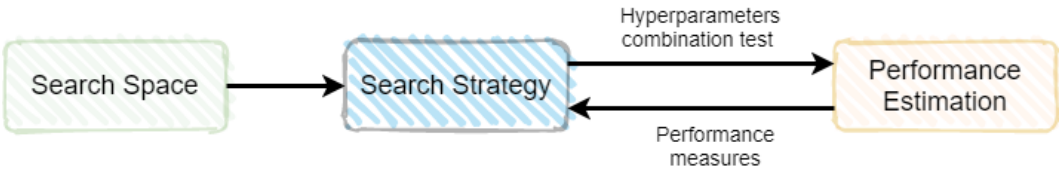


Figure 5.3 Illustration of the Neural Architecture Search (NAS) concept

As previously mentioned, a great amount of the existing work in deep learning is focused on 2D CNN and thus, the scarce NAS work is mainly focused on image processing with the use of 2D CNN. There are also some research works that have started to emerge in topic such as semantic segmentation, deep reinforcement learning and machine translation [177] [178] [179].

The research documented in this chapter started from the research initiative document in the chapter 8.7 Architecture Optimization, use case “Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children, where it was necessary implement and optimize a machine learning model for the classification of a 1D signal.

The CNNs were a promising solution for the research goals, but there was the need to find the best architecture to solve the classification problem, in this case, classify PQRST cardiac complexes (or in a less clinical way, heartbeats). The literature review has revealed some strategies, such as the application of trial-and-error, greedy, random and Bayesian approaches, but to the best of our knowledge, there were no libraries or frameworks able to process the 1D CNN architecture optimisation, neither comparison between approaches which could help the research team on the work that needed to be done.

To address this gap, the team has developed NAS processes for 1D CNN architectures and perform the comparison between the five different search strategies: greedy, random, Bayesian, hyperband and genetic, in order to understand which strategy best fits the requirement of achieving an optimised architecture.

The following sections offers a detailed view on how the five strategies behave, and it also shows the process of finding the best 1D CNN architecture based on performance stability, running time and computational resources. This information is very useful, allowing researchers and developers to make better decisions when facing the need to use a NAS process for a 1D CNN problem.

In resume, there is a huge potential in the use of 1D CNN but, currently, researchers seem to avoid 1D CNN due to lack of expertise or use unoptimized 1D CNN due to lack of optimisation tools. This research intended to overcome these limitations, presenting the implementation of NAS approaches for 1D CNNs, as well as a search strategies comparison analysis, which certainly will benefit the community on 1D data modelling with the use of optimised 1D CNNs.

5.1.2 1D CNN NAS Testing and Measurements

The CHAPTER 3 - Methodology(ies) and in particular the subsection titled 3.3 Optimization Algorithms Benchmark, describes the methodology used in the current work to test and evaluate the NAS approaches for 1D CNN. To ensure a comprehensive, unbiased, and careful evaluation, it was decided to use a tested methodology, and the one presented was undoubtedly the most suitable. The testing and measurements process is covered in detail in this chapter, including all steps and choices.

5.1.2.1 Reason for Comparison

Regardless of the application area, data on the 1D shape is frequently found when working with sensors, smart sensors, the Internet of Things (IOT), and other collection data instruments. For instance, in the healthcare area, one of the most common diagnostic methods is the electrocardiogram (ECG) which produces 1D data. Specifically, in a real case research on the medical area involving ECG medical exams, it was necessary to develop a 1D CNN able to process the ECG waves for a real classification problem (a supervised machine learning problem), supported by real patient data. This research is documented in the CHAPTER 8 - Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children.

The performance measured by the accuracy of the network was crucial for the success of the research, and the need for defining the optimal 1D CNN architecture for the problem arose. After analysing related works, it was possible to understand that the problem could not be solved with a mathematical formula or other similar approach, and thus it was required to perform an optimal hyperparameter combination search. The concept of optimal hyperparameter combination search for neural networks is named Neural Architecture Search, commonly known as NAS.

The literature review also revealed that NAS could be developed following some alternative strategies, but no comparisons between alternatives were found and it was not possible to understand which options would lead to the best solution, namely find the optimal hyperparameter combination for the 1D CNN for a given problem. It is also important to highlight that, even though it was possible to find some work related to the 2D CNN optimisation architecture search, it was not possible to find NAS approaches for 1D CNNs [180].

The motivation for creating this comparison was to understand which NAS approach would be able to find the best model architecture, or technically, the set of hyperparameters which perform best for a given problem. The ECG classification was the use case chosen as input to support the comparison. Furthermore, the motivation can be broken down into the following aspects (by order of importance):

1. Model architecture with the best accuracy: the performance of neural networks can be measured through several different metrics and according to the machine learning problem typology. Some examples of metrics are root mean squared error, mean squared error or adjusted r squared (R^2) for regression problems, or the accuracy, precision, or F-Score for classification problems. In the present use case, a classification problem, the best metric is the classification accuracy. The accuracy is calculated following the following equation.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negative}}{\text{True Positives} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \quad (5.1)$$

2. Algorithm execution time: the critical success factor was the 1D CNN accuracy, and there was no need to define the neural network architecture in real-time or in a very short period. Still, it was a requirement to find the optimal architecture on a realistic timeframe, which, has been defined, according to the team requirements, as “good” if the total running time is below 24 hours and “excessive” if the total running time is above 32 hours, considering the problem complexity. 24 hours usually corresponds to a working day. It is considered as “excessive” (above 32 hours) since it requires waiting for the results for more than two working days.
3. Computational resource’s consumption: the democratisation of machine learning tools and methods is only possible if they can be executed on accessible/affordable computational architectures. In this initiative, it is considered that the NAS should be executed on a relatively cost-accessible infrastructure for any research centre in most of the countries in the world.

5.1.2.2 Test Set

As previously referred, the need to build an optimised model based on 1D CNN arose from a real-world use case research scenario, where there was the need of classifying ECG waves (a supervised machine learning problem). The dataset used as the test set corresponds to the real-world use case dataset.

Briefly, the ECG data was collected by the Generation XXI project, a previously established prospective population-based birth cohort from Porto, Portugal, that aimed to characterise prenatal and postnatal development throughout childhood, adolescence, and adulthood. This project followed up a total of 8647 children born between April 2005 and August 2006 and gathered several ECGs among the cohort population at the 10 years' evaluation.

The data was extracted from an ECG equipment that was assigned to the Generation XXI project, in a proprietary vendor XML format. Data was processed by the research team which has performed several machine learning tasks to prepare and achieve a final dataset.

The preparation phase has included a data understanding phase, where the XML detail information was analysed, and relevant data (future features) were identified. The following information was extracted from the XML format: ECG collection date; individual related data, namely the identification (encoded), birthdate, weight, height, gender, ECG signal (Lead I) and ECG report. Regarding the ECG signal, the records presented a sampling rate of 500 samples/s, with a ten-second duration, which results in a total of 5000 measures for each record. The ECG was collected in a raw format, which means noise filtering was not performed. The data, which included records from 5646 children, was also inspected to find duplicates, records with missing and/or abnormal signal data, incorrect birth date or collection time, invalid weight and/or height values. The data understanding phase was completed with an extensive statistical and data analysis stage.

The data preparation phase included the creation of derivative features, elimination of records that presented cardiac abnormalities (identified by the ECG report feature in the form of text), the ECG signal filtering with a Finite Impulse Response (FIR) and a band-pass filter, and the decomposing of each of the 10.000 millisecond signal in beats, supported by the identification of R-peaks followed by the beat extraction.

More details about the test set can be found in the CHAPTER 8 - Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children section.

The process resulted in the extraction of a total of 14.308 complete beats, with a 600 ms duration, composed by 300 samples (one sample by each 2 ms).

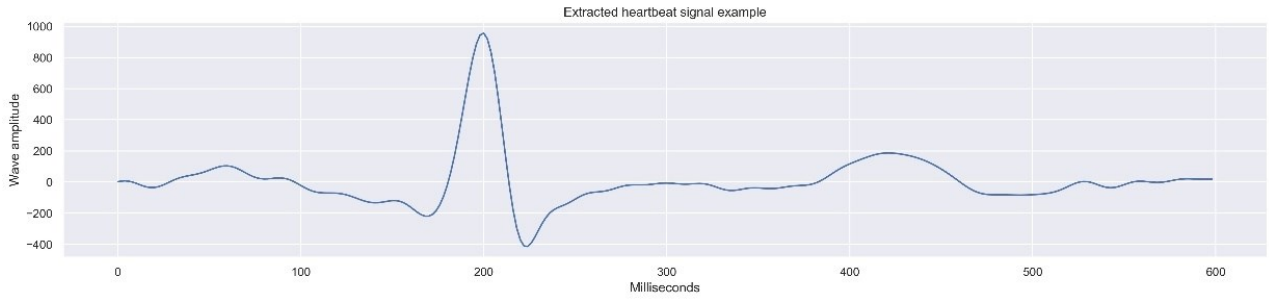


Figure 5.4 1D data example (heartbeat signal)

In light of the NAS testing and measurements, it is important to emphasise that all performed tests were based on the same testing set, exactly on the same conditions (topic further discusses in the next section), following the same training process, including record order.

5.1.2.3 Perform the Experiments

As discussed in the methodology section (3.3 Optimization Algorithms Benchmark), the experiments rely significantly on the environmental factors, which include computational settings, but also the programmer's skills and the programmatic environment.

In order to minimise noise or distortion on the experiments, the same machine with the same settings were used. The research team also "froze" any software update that could be performed automatically, turned-off schedule works and limited the access to the environment only for the evaluation purposes. To avoid any influence of the programmer's skills in the NAS processes and strategies, as well as the implementation of the comparison, only a single programmer worked on the entire development. The CNN's model training was performed in the same way, based on the Keras framework, assuring the focus on the NAS search strategies comparison, and how strategies combine the hypermeters. Keras framework is an open-source software, written in Python, developed by Google for implementing neural networks [181]

A highlight for the fact that the combinations of training hyperparameters influence in fact the training. Finally, the same implementation language was used, in this case, Python 3.7.

The computational infrastructure that supported the experiments had the following settings:

- OS: Ubuntu 18.04;
- Disk: 500 GB SSD;
- Sys Mem: 32Gb;
- CPUs: AMD Ryzen 7 3600X;
- GPUs: NVIDIA GTX 1080 TI 11GB + RTX 2060 SUPER 8GB.

This configuration was chosen in accordance with the guidelines outlined in the previous phase, "Reason for the Comparison". These guidelines state that NAS should be run on a computational architecture that is both accessible and affordable. Despite using a relative basic hardware setup, this infrastructure is very affordable and has an interesting computational capacity.

For the experiments, considering the relevant algorithm metrics to be assessed, the following measures were monitored:

- Model accuracy: In this current use case, the research team has defined classification accuracy as the suitable performance metric. The accuracy was measured in the three sub-datasets extracted from the initial dataset. The sub-datasets were named as training, validation and testing sets, and the split from the original set was 70%, 20%, and 10% for each set, respectively. The goal was to achieve a model that best generalises the learning on data and thus, a solution where accuracy values were homogeneous and consistent was considered as a good solution, instead of “simply” the best accuracy values of any of the sub-datasets.
- Running time: as previously defined, there was a limit (defined by the research team) for the execution time capped by 32 hours and a desirable run time lower than 24 hours. For that reason, the total running time was measured.
- Resource’s consumption: the benchmark experiments and NAS strategies usually consume a considerable amount of CPU and GPU resources, especially GPU usage and memory. So, it was necessary to be constantly monitoring those metrics.

In summary, the experiments were measured on the quality of the solution (variant unknown solution) on the accuracy of the three partial data sets; and on the efficiency through execution time and resource consumption.

As described on previous sections, it was not possible to find NAS frameworks or tools that processed 1D CNNs and those were necessary to develop / adapt / extend new processes / methods for 1D CNN NAS. The NAS strategies (depicted in Figure 5.5) considered in the current comparison followed two main approaches, namely:

- Approach A: Custom processes based on AutoKeras, an AutoML library that provides some tools for hyperparameter optimisation. The developments have included four different strategies: greedy, random, Bayesian and hyperband [182] [183].

- Approach B: a new development based on the Deep Neural Network Evolution (DEvol) library. DEvol presents a basic proof of concept for genetic architecture search in imaging data [184].

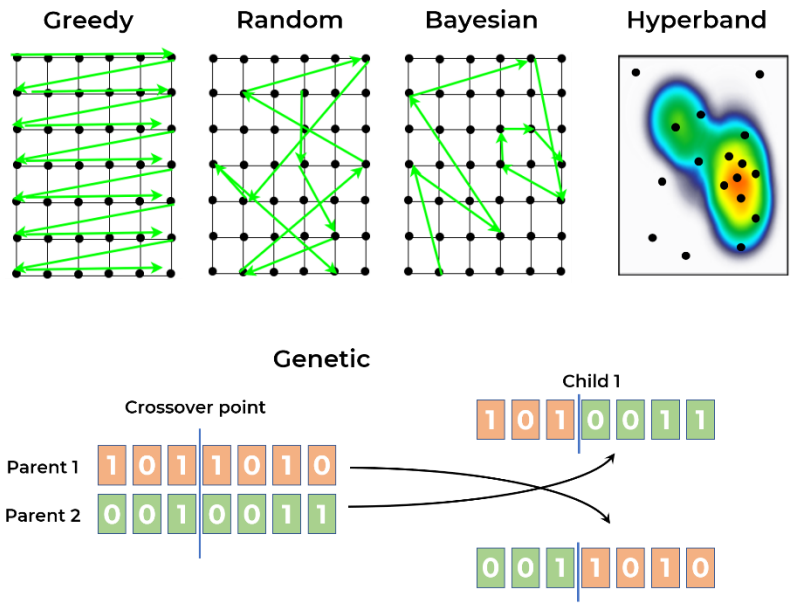


Figure 5.5 Illustration of the NAS strategies applied in experiments

The experimental testing design was the following:

- Dataset
  - 14.308 records (complete beats)
  - Split: 70/20/10 (training, validation, and testing)
- 25 experiments for each scenario
- Scenarios:
  - Approach A
    - A1: Greedy
    - A2: Bayesian
    - A3: Hyperband
    - A4: Random
  - Approach B: Genetic

The greedy approach is an algorithm that can find the best solution in a limited search space because it tries out all possible combinations. In a non-limited space or where not all possible solutions can be tested, the algorithm tries the possible combinations within a defined timeframe or number of attempts. The search strategy consists of trying all possible values for each variable and then changing only one in each iteration. The random approach is similar, but in this case all variables are chosen randomly in each iteration. Although trial-and-error, random and greedy strategies are known to be the most common approaches, in more recent 1D CNN research (as previously commented in the introduction section), there is no orientation / direction in the search process to find the best solution.

Bayesian search considers the last combination of variables and the corresponding score / result when defining the parameters for the next iteration. The algorithm optimises the iteration possibilities, i.e., it tries to search in promising areas of the search space, increasing the probability of achieving the best combination of variables.

The concept of the hyperband search consists of eliminating or early stopping the solutions that don't perform well and those that don't "seem promising", to reduce computational resources. It starts with a small exploitation in each of the possible combinations, and if the performance of this attempt does not perform well, the combination is discarded for the next iteration of the algorithm. This cycle repeats itself until there is only one combination left.

The genetic algorithm is a search algorithm inspired by the theory of natural evolution. The concepts reflect the process of natural selection where the best individuals in a population are selected for reproduction, and the next generation is expected to inherit the best characteristics of the best parents, contributing to an improved next generation. In this particular case, the characteristics of the individual are the hyperparameters choices/combinations, specifically the ones described in the next paragraphs. It is an iterative process that seeks to achieve a final population made up of the fittest individuals, that hopefully includes the best possible solution for a given problem.

Each of the strategies is described in greater detail in chapter 4.1 Architecture Optimization .

All the previous scenarios perform a search for the best CNN hyperparameters, including the structural and training hyperparameters, following each of the described strategies. The considered structural hyperparameters were the number of layers (convolutional, pooling, and fully connected layers) that compose the CNN structure, as well as the number of filters, kernel sizes, padding, stride, activation functions, normalisation techniques, dropout percentages in each layer and number of neurons (for the fully connected layer). Regarding the training parameters, the search was performed on the network optimisation function, learning rate, batch size, weights initialisation, bias initialisation, and momentum.

For a fair comparison between approaches and considering the NAS goals, the search settings were defined in a way that guaranteed equal opportunities for all scenarios. NAS configuration settings: epochs: 200; max trials: 400; generations number: 20; population size: 20.

The application of the five scenarios and the corresponding experiments have resulted in the best 1D CNN architecture for the problem (accuracy optimization), which is presented in Figure 5.6.

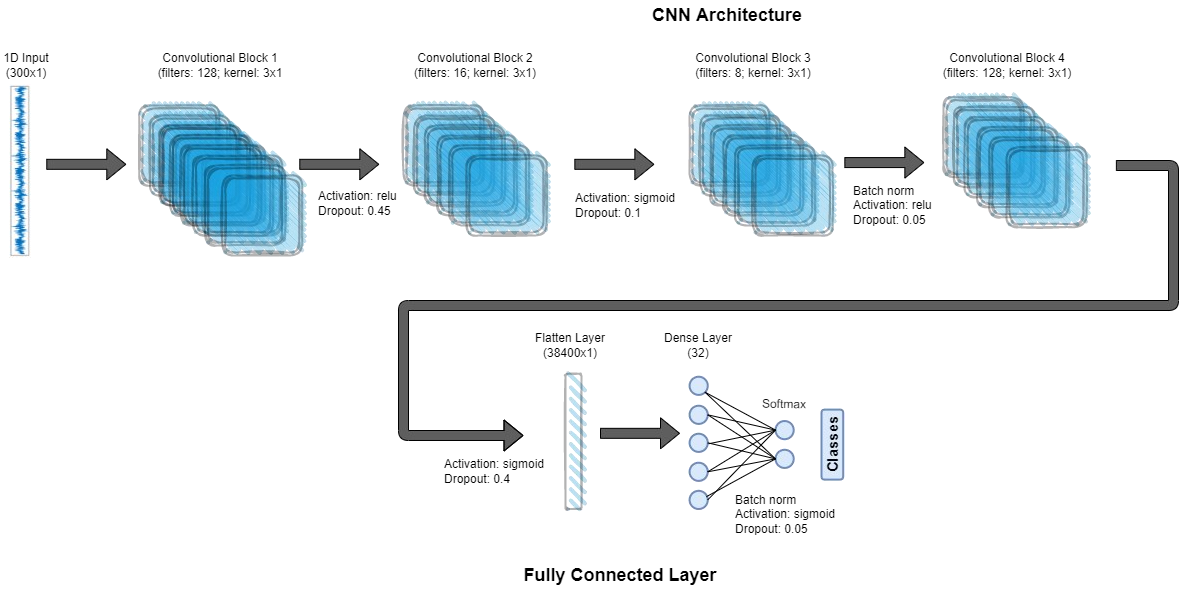


Figure 5.6 Optimized 1D CNN architecture produced by the NAS process.

5.1.2.4 Analyse & Report

Considering the reason for the comparison and the important aspects to evaluate, it was decided to create a summary metric table (e.g.: Table 5-1) with the accuracy outputs for each scenario from the 25 experiments.

To evaluate the results achieved by the models, the table contains the accuracy measured in the three different data sets, namely training, validation and testing (70/20/10 % split), as well as the most consistent model. The training set was used to minimise an error function in the modelling phase. The results and the quality of the solution are measured by analysing the performance of the validation and test subsets. The most consistent model is obtained by analysing the 25 experiments for each scenario and measuring them from the difference resulting from the accuracy of the validation and test datasets in each iteration. A stable model shows better generalisation resulting in more consistent results, which means less difference between the two accuracy values.

Considering the non-deterministic nature of a 1D CNN implementation and the search strategies applied on the NAS, the research team has considered the variability / dispersion in the performance results to be relevant, and thus box plots were developed to analyse output stability. This dispersion analysis allows avoiding a noisy or distorted evaluation due to a "lucky shot" on any of the scenarios.

A summary table, the most consistent model, and the accuracy box plot are presented in the following tables and figures, for each of the scenarios. The presented accuracy is based on the classification performance, (multi-class accuracy) which is defined as the average number of correct predictions. For this measurement, a confusion matrix was used since it presents both predicted and actual/true class summary, highlighting the sum of true and false predictions.

Table 5-1 Summary table for Approach A1 - Greedy accuracy (in %)

Approach A1 – Greedy

Accuracy in %	**Train (%)**	**Validation (%)**	**Test (%)**
Best	0.7876	0.8113	0.8057
Worst	0.7616	0.7896	0.7442
Mean	0.7784	0.8007	0.7737
Std dev	0.0070	0.0049	0.0144
Most consistent model		0.7934	0.7903

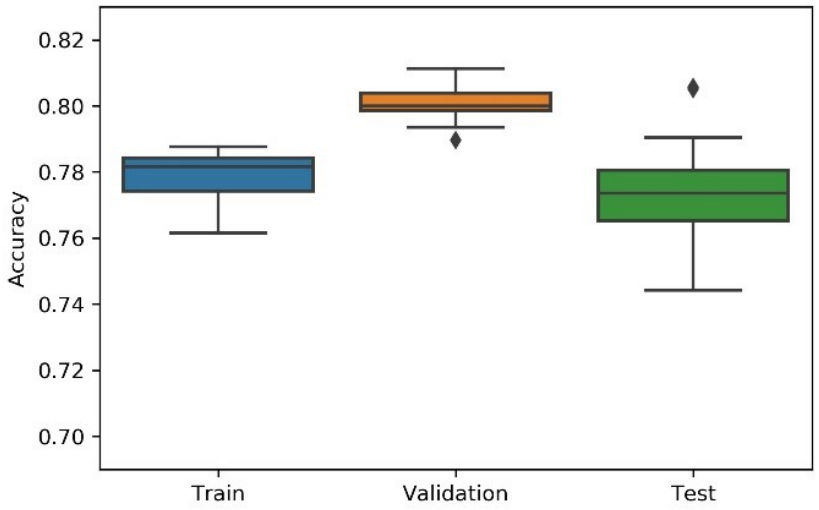


Figure 5.7 Experiment's performance for Approach A1 (Greedy) accuracy (in %)

Table 5-2 Summary table for Approach A2 - Bayesian accuracy (in %)

Approach A2 – Bayesian

Accuracy in %	Train (%)	Validation (%)	Test (%)
Best	0.7969	0.8050	0.7694
Worst	0.7460	0.7938	0.7093
Mean	0.7711	0.7971	0.7353
Std dev	0.0150	0.0029	0.0178
Most consistent model		0.7945	0.7624

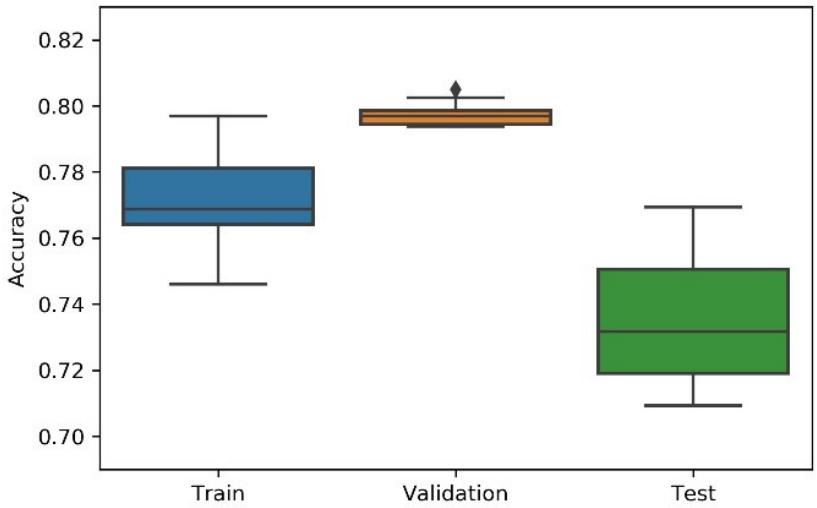


Figure 5.8 Experiment's performance for Approach A2 (Bayesian) accuracy (in %)

Table 5-3 Summary table for Approach A3 – Hyperband accuracy (in %)

Approach A3 – Hyperband

Accuracy in %	Train (%)	Validation (%)	Test (%)
Best	0.7580	0.7948	0.8001
Worst	0.7190	0.7550	0.6925
Mean	0.7328	0.7587	0.7332
Std dev	0.0072	0.0081	0.0231
Most consistent model		0.7613	0.7582

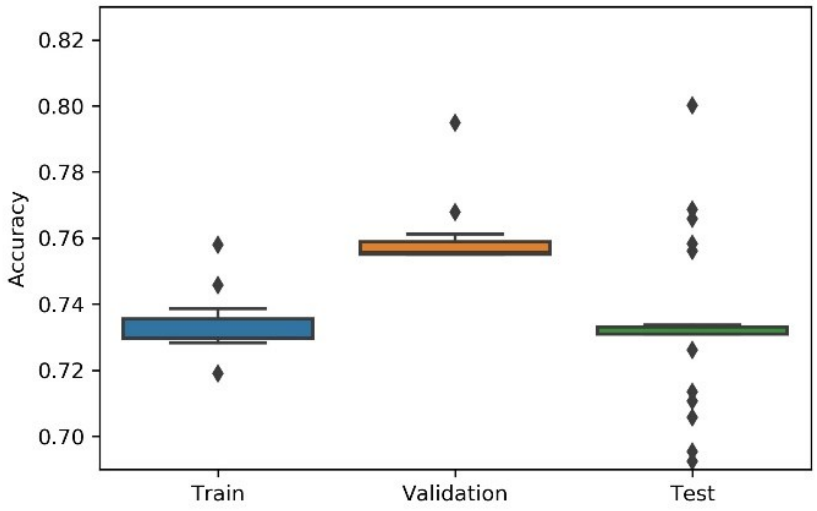


Figure 5.9 Experiment's performance for Approach A3 (Hyperband) accuracy (in %)

Table 5-4 Summary table for Approach A4 – Random accuracy (in %)

Approach A4 – Random

Accuracy in %	Train (%)	Validation (%)	Test (%)
Best	0.7948	0.8022	0.8106
Worst	0.7544	0.7903	0.7345
Mean	0.7709	0.7955	0.7597
Std dev	0.0127	0.0025	0.0166
Most consistent model		0.7983	0.7932

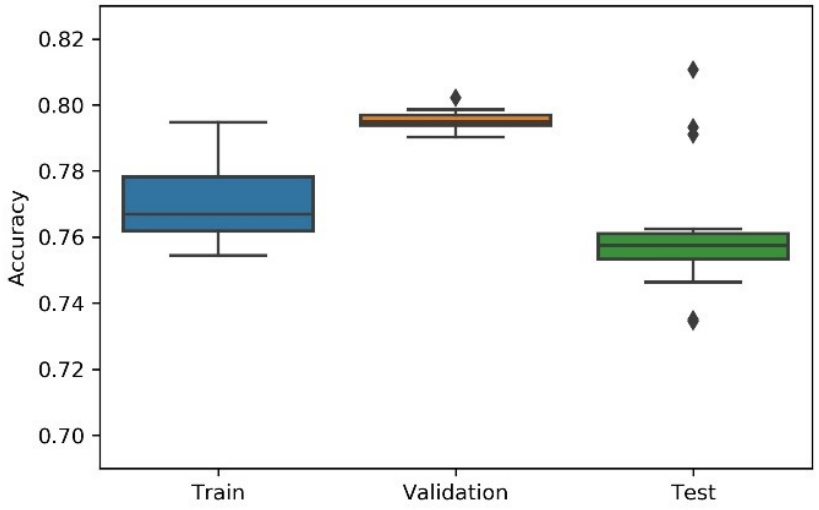


Figure 5.10 Experiment's performance for Approach A4 (Random) accuracy (in %)

Table 5-5 Summary table for Approach B - Genetic accuracy (in %)

Approach B – Genetic

Accuracy in %	Train (%)	Validation (%)	Test (%)
Best	0.8240	0.8067	0.7980
Worst	0.7716	0.7882	0.7491
Mean	0.7887	0.7964	0.7799
Std dev	0.0116	0.0040	0.0124
Most consistent model		0.7960	0.7980

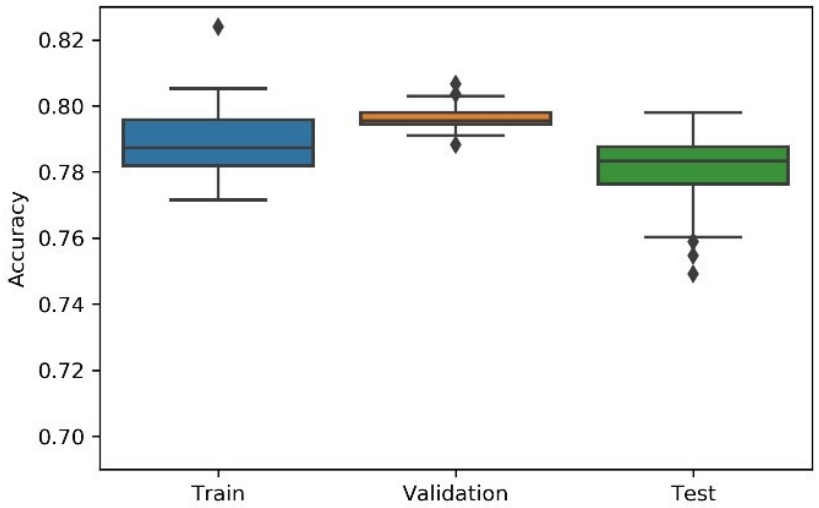


Figure 5.11 Experiment's performance for Approach B (Genetic) accuracy (in %)

The second measure defined as relevant for monitoring was the execution time. The execution time was obtained by measuring the execution time of the algorithm from the initial start of the algorithm to the result consisting of the best network found. Recall for the NAS settings, namely 200 epochs, 400 max trials, 20 generations and a population size of 20 samples.

Table 5-6 shows a summary of the execution times for each scenario.

Table 5-6 Experiments measures: running time

Scenario A1 – Greedy

Run time	Time (h:m:s)
Best	08:08:14
Worst	30:55:25
Mean	13:48:27
Std dev	04:56:32

Scenario A2 - Bayesian

Run time	Time (h:m:s)
Best	08:25:00
Worst	28:51:00
Mean	16:55:57
Std dev	04:55:31

Scenario A3 – Hyperband

Run time	Time (h:m:s)
Best	00:57:00
Worst	1:05:00
Mean	01:01:07
Std dev	00:02:27

Scenario A4 - Random

Run time	Time (h:m:s)
Best	11:19:00
Worst	12:13:00
Mean	11:37:04
Std dev	00:09:41

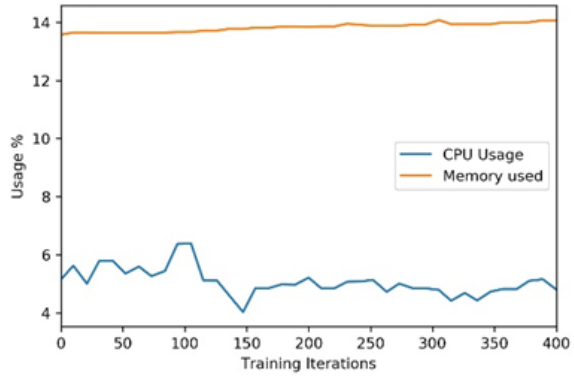
Scenario B - Genetic

Run time	Time (h:m:s)
Best	09:21:00
Worst	28:20:00
Mean	14:58:40
Std dev	04:46:36

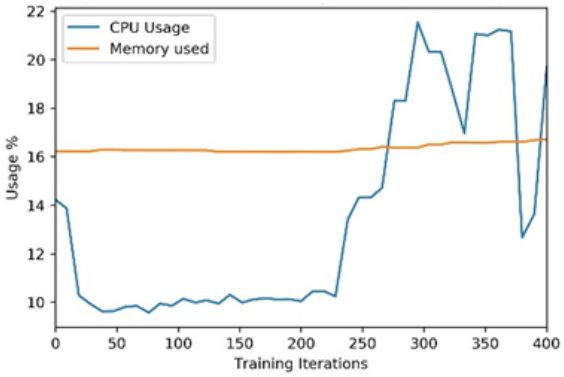
The current comparison process includes the monitoring of computational resource consumption. In Table 5-7 it is possible to find CPU and GPU resource consumption, including GPU memory, for each scenario. The resource consumption was analysed for each scenario considering all 25 experiments. By analysing the CPU and GPU monitoring results on different experiments, similar results were obtained, so it was decided to provide a resource monitoring sample for each scenario. Based on the CPU consumption, the vertical axis in the figures shows the percentage of the CPU used.

Table 5-7 Experiment's measure: CPU consumption

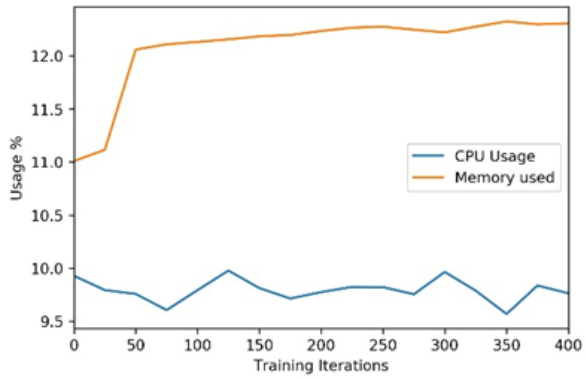
Scenario A1 - Greedy



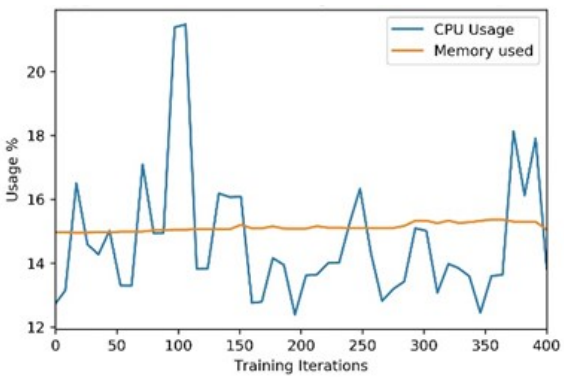
Scenario A2 – Bayesian



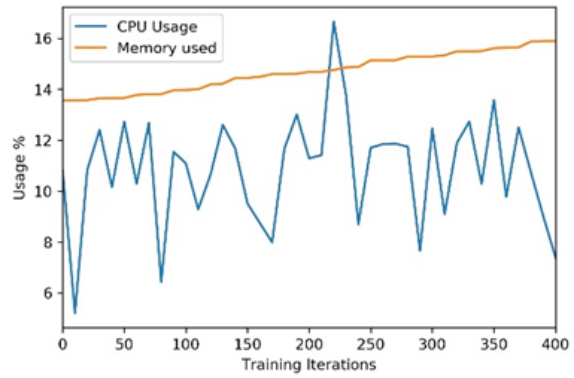
Scenario A3 – Hyperband



Scenario A4 – Random



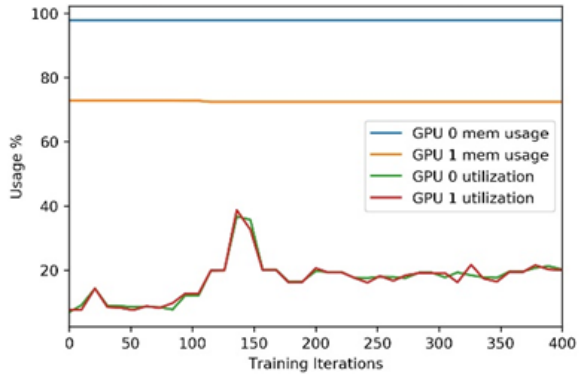
Scenario B – Genetic



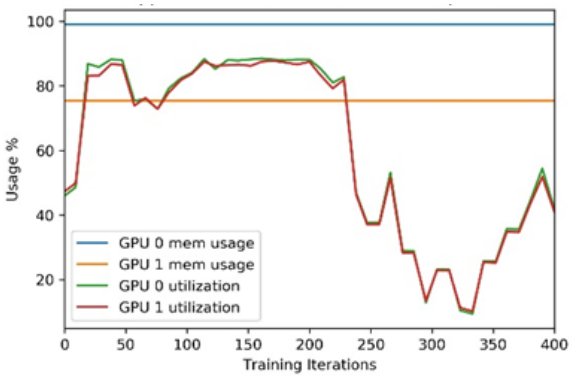
Regarding GPU resource consumption, the test environment infrastructure included two GPUs. Table 5-8 presents the percentage of GPU usage and GPU memory usage.

Table 5-8 Experiment's measure: GPU consumptions

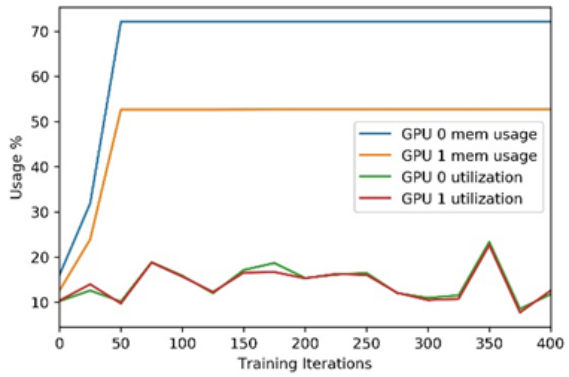
Scenario A1 - Greedy



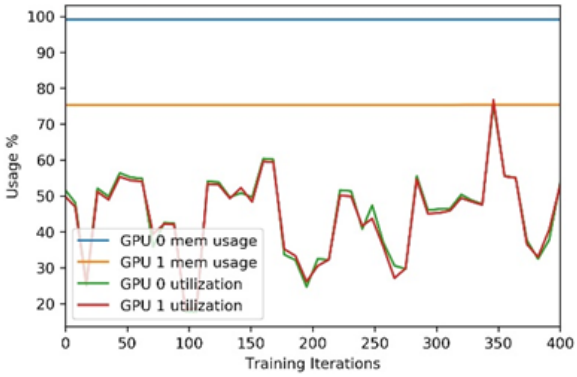
Scenario A2 – Bayesian



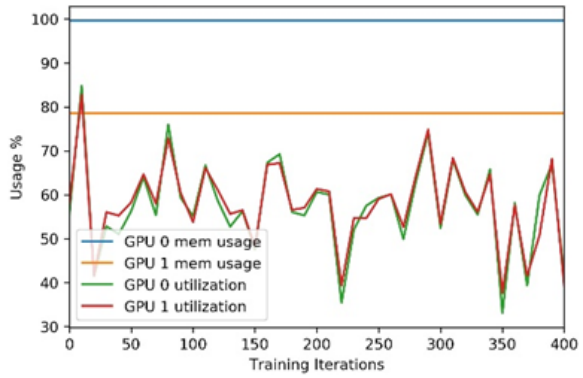
Scenario A3 – Hyperband



Scenario A4 – Random



Scenario B – Genetic



The plots were created using the data generated by the experiments and managed by the open-source library ClearML - Auto-Magical Suite of tools to streamline your Machine Learning workflow Experiment Manager, ML-Ops and Data-Management [185].

The experiments and the experiments outcomes can me resume by the following Table 5-9:

Table 5-9 Experiments Testing and Measurement overview

Measures	Approach A1	Approach A2	Approach A3	Approach A4	Approach B
Most consistent model	**0.7777/0.8001/0.7750**	0.7688/0.7945/0.7624	0.7458/0.7581/0.7687	0.7710/0.7941/0.7624	**0.7985/0.7959/0.7918**
Mean run time	13:48:27	16:55:57	01:01:07	11:37:04	14:58:40
Worst run time	30:55:25	28:51:00	01:05:00	12:13:00	28:20:00
CPU (consumption %)	<7%	<22%	<10%	<22%	<18%
Memory 32GB (%)	<145%	<17%	<13%	<16%	<16%
GPU 0 Memory (%)	99%	99%	<70%	99%	99%
GPU 1 Memory (%)	76%	76%	<53%	76%	76%
GPU 0 Utilization (%)	<39%	<89%	<23%	<78%	<86%
GPU 1 Utilization (%)	<39%	<89%	<23%	<78%	<86%

Considering the experimental tests and the accuracy metric, it is possible to verify that:

- The best accuracy in the validation sub-dataset was achieved by Approach A1 (greedy) and Approach A4 (random) achieved a value of 0.8106 (%) in the testing sub-dataset.
- By looking more closely at the full tables of results, it is possible to verify that the value obtained on Approach A4 (random) for the test sub-dataset was a “lucky shot”, and when observing the distribution of results through the box plot, we confirm that this value is clearly an outlier.
- Approach A3 (hyperband) presents the lowest values for the worst outcomes of all subsets of data. It also presents the lowest scores when considering the mean results. At first glance, this approach seems to have the worst accuracy ratings, but in order to understand these values and to draw a conclusion, it is important to understand how the hyperband approach behaves. We recall the description of the hyperband strategy (on a previous section) which relies on an early stopping strategy for resource allocation and run time optimisation.
- The approach that achieves the best mean accuracy is Approach B (genetic) on the test sub-dataset and Approach A1 (greedy) on validation set.
- Regarding standard deviation, Approach A3 presents the highest values for the validation and test sub-dataset. The lowest values are presented by A4 and B respectively, for validation and testing sets.
- Analysing the boxplots, Approach A3 stands out as the one presenting the most narrows interquartile range, and the one with the most outliers. Both approaches A1 and B show consistency between the three partial data sets, which show box plot shapes with similar accuracy. Two differences stand out among these approaches: i) the three subsets of data have similar shapes, but the training, and mainly the testing have higher scores in approach B; ii) in the test subset, A1 represents higher values as outliers, with B representing outliers at the lowest values. A2 is the approach with the lowest consistency, as well as the smallest values on the accuracy distribution in the testing box plot.
- The results of the box plot analysis are confirmed by comparing the “most consistent model” for each approach. Approach B shows the smallest deviation in the validation and test subsets with accuracies of 0.7960 / 0.7980 (%), followed by Approach A1 with 0.7934 / 0.7903 (%). The highest deviation is presented by A2, followed closely by A4.

Considering the running time analysis, it is possible to verify:

- With an execution time (at least) 8 times shorter, the A3 approach performs better than all other variants. This can easily be explained by the approach behaviour (hyperband), previously described.
- Approaches A1, A2 and B showed similar values. These results were expected because all approaches perform experiments similarly. The biggest difference lies in the search design, whereby A1 is “dummy”, A2 offers an “oriented” search and B defines new possible solutions based on genetic algorithms.
- Scenario A4 shows a reduced variation considering the total running time.
- All approaches carried out in an acceptable time, which, given the definition of the study, was 32 hours. Most were able to run under the desirable 24 hour timeframe.

Regarding the computational resource consumption:

- In all cases, the CPU usage was below 22% and the system memory usage was below 16%. (Available RAM: 32 GB). It is also important to note that the experiments were configured to use GPUs for CNN processing to optimise convolution operations. Hence, it was expected that it would not have high CPU or system memory consumption.
- All approaches used the same training process that uses all the available GPU memory. For this reason, the approaches have a similar memory consumption (note that some of the GPU-1 memory is reserved for system tasks). The exception can be found in the A3 scenario, which starts with a few epochs in every training session and then reduces the number of trained models.
- The utilisation of both GPUs is identical in all scenarios, showing a parallel consumption. This behaviour was expected because the training was configured to distribute processing among the GPU units.
- The GPU utilisation ranges from 10% to 90% in approaches A2, A4 and B and is significantly lower in A1 and A3.

5.1.3 Conclusions

This chapter starts by introducing the context with an overview of deep learning algorithms and more specifically on CNNs, including 1D, 2D, and 3D convolution processing alternatives. Following, is provided the problem characterisation that consists in optimising a 1D CNN architecture (best options for training and structural hyperparameters) and introduces the concept of NAS as a solution. To optimally achieve the goal of optimising a 1D CNN, and since the existing literature does not provide, for now, an optimal way for reaching an optimised 1D CNN (the research gap), five different search approaches / procedures have been implemented, tested, and measured.

The approaches include five NAS strategies, namely greedy, random, Bayesian, hyperband and genetic. For the experiments and analysis, several dimensions were considered, namely accuracy (with different perspectives), total execution time and resources consumption.

The best performance was measured by the accuracy on the validation and testing sets, as well as the models' consistency and dispersion of the results. Though approach A1 (greedy) has reached good results on the accuracy, approach B (genetic) has the best outcomes when analysing the model's consistency and the dispersion of the results (best observed on the presented box plots).

Regarding efficiency, measured by the computer system resources usage and total execution time, all algorithms had a very acceptable behaviour with decreased consumptions of CPU and system memory. Regarding GPUs, the algorithms have taken full advantage of their capacities for CNN training, with high use of the available resources. Approaches A1 and A3 were a little less demanding. This outcome was expected considering that one of the 1D CNN advantages is a reasonable demand of computational resources.

Total running time was acceptable in all approaches. The average run time was below 24 hours which the study defined as a good total execution time. Even the worst cases exceed this limit on a maximum of 7 hours which was still acceptable. A final note for the approach A3 which stands out with the need of only 1 hour processing, which is 10 times less than any other. This approach is a good choice when facing a short time limitation and a less performant network is acceptable.

Considering the initiative definitions and purpose, the research team concluded that the NAS strategy that best performed was the approach B, the genetic approach, which uses affordable resources (like the other alternatives) and has produced the most performant architecture for the present use case challenge.

The optimisation of neural networks in general, and mainly for 1D CNNs, is critical for the extraction of data's true value and meaning. Nevertheless, the literature review showed that, even in the most recent works, the common approaches fall back to trial-and-error methods and grid searches with limited parameters. These current approaches present as drawback the non-effective optimisation of networks, meaning potential wasted data value. The NAS approach and the resources presented, intended to contribute to the model's optimisation, and consequently to improved research results. The use of NAS presents the disadvantage of adding one more phase to the already complex machine learning pipeline. In our opinion, the benefits greatly outweigh the costs.

The research team is confident that this initiative, besides being critical for the use case documented in CHAPTER 8 - Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children, will provide valuable knowledge and resources to future 1D CNN researchers in the domain of sensors and other 1D data sources.

5.2 Developments and Implementations

In the thesis context, namely in the challenges and works described in this document, the research team was confronted with some limitations and/or a lack of tool (e.g.: 1D CNNs optimization). In order to go forward with the stated aims and goals, new tools and functionalities have been created. A list of the most pertinent works is provided below:

- Extension of BioSppy open source library, where improvements were made to enrichen signal processing, specifically in the area of the electrocardiogram. These developments were essential for the research documented in CHAPTER 8 - Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children
- The expansion of the Auto-Sklearn open source library, with optimization methods for some regression algorithms (e.g.: XGB, LightGBM). This optimization methods were used in the Architecture Optimization phase, in the Applied Data Science: Length of Stay in Neonatal Intensive Care Units use case Architecture Optimization
- Improvements to AutoKeras open source library, providing developments for 1D CNN NAS processing. These developments were both used in Neural Architecture Search for (1D) CNNs and in Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children
- The addition of 1D CNN NAS to DEvol - Deep Neural Network Evolution open source library, which, similar to the previous point was used in Neural Architecture Search for (1D) CNNs and in Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children
- Adaption of Grad-Cam techniques to 1D processing. This explainability technique application is document in 8.9 Explanation (Analyse of Cardiovascular Changes Due to Obesity in Children use case)

In addition to being useful for this thesis, we anticipate that these developments will be advantageous for upcoming researchers, regardless of the business area in analysis. Data science is a broad field, and every project or research has its own distinct characteristics, but machine learning models and architectures optimization also face similar difficulties.

The research team is anticipating that upcoming initiatives will run into challenges with demands similar to those this thesis faced but this time, the solutions here provided should assist in overcoming those challenges.

CHAPTER 6 - Applied Data Science: Child's Target Height Prediction

This chapter is the first of three chapters on applied data science in children healthcare. The chapters go into detail about real use cases that were developed using the methodology that was previously suggested for the thesis, sub-chapter CHAPTER 1 - 3.2Thesis Framework.

The main objective of the PhD thesis is the application of data science to contribute to increasing the health and well-being of our population, extracting value and/or knowledge that is hidden in already existing data. This value/knowledge can be materialised in new solutions that create or improve existing methods and/or answers to scientific questions that were unknown until today.

Considering that the data science clinical field is a vast area it was decided to focus on children, a population segment which usually get less attention from research, since they (fortunately) don't have a high prevalence of pre-existing and/or diagnosed diseases, the more immediate health challenges are observed in elderly people, among other reasons previous approached in the Data Science in Children Healthcare section. However, it is important to develop research work in these young ages as well, as they will contribute earlier to the health and well-being of the overall population presenting better outcomes (e.g.: early identification of diseases and their treatment at more favourable stages, or correction of physical issues in the growth phase that would manifest their effects over time).

This particular research work, documented in this chapter, is a contribution for the improvement of healthcare in children and in society in general, focused on the growth assessment field. The goal of this study is to forecast children's "target height" as adults in order to improve growth assessments and provide more individualized healthcare. The existing literature describes some existing prediction methods, based on longitudinal population studies and statistical techniques, which, with few information resources, are able to produce good results.

The challenge of this study was adopting a novel machine learning-based strategy to forecast the target height and (eventually) enhance the accuracy of the current height prediction models. The study's objectives were met.

In terms of height prediction, the extreme gradient boosting regression (XGB) and light gradient boosting machine regression (LightGBM) methods produced noticeably improved results. Paediatricians and other healthcare practitioners can usefully apply the created model to assess growth in their patients.

This research work was published in the Applied Sciences, a journal classified, for instance by ISCTE-IUL, in the Q1 Quartile. The paper "Child's target height prediction evolution" has already received some citations, what makes us very proud.

6.1 Introduction

Children's growth assessments are a very important aspect of health status monitoring, for example, to identify development deviations, and diseases or interventions/treatments impacts (e.g.: obesity care or stunting). Monitoring the children's growth helps to ensure children develop in the best possible way [186] [187] [188] [189].

But what exactly is growth and growth assessment? Growth is a complex interaction between heredity and environment. Growth refers mainly to changes in magnitude, increments in the size of organs, increases in the tissue thickness, or changes in the size of individuals as a whole [190].

The study of growth, in a more scientific definition, is focused on monitoring individuals that have not reached the "mature" stage, where full growth potential should be realized [191].

Knowing the importance of growth assessment several growth standards have been developed in the past decades, for instance (the most popular) NCHS/WHO, CDC 2000, Tanner standards, or Harvard standards [186] [192].

In 1978, the NCHS—National Center for Health Statistics growth reference (the most-used reference until then) was replaced after a recommendation of The World Health Assembly in 1994, which considered the reference did not represent early childhood growth [193].

The NCHS reference was based on a longitudinal study of children from a single community in the USA, which were measured every three months [194] [195].

The (new) WHO-World Health Organization Standards were developed using data collected on the WHO Multicentre Growth Reference Study (which combines a longitudinal study with a cross-sectional study) between 1997 and 2003. Data included about 8500 children from South America (Brazil), Africa (Ghana), Asia (India, Oman), Europe (Norway), and North America (USA) [195] [196] [197]

A global survey was conducted in 178 countries to understand the use and the recognized importance of the new standards. The importance of this work was recognized after huge scrutiny and a wide adoption for more than 125 countries just five years after the work was released (April 2006). Child growth charts are currently one of the most used tools for assessing children's wellbeing and, allow the physiological needs for growth to be followed and establishing whether the expected development of the child would be met [186] [196].

However, it's important to verify that the main reason for non-WHO Standards adoption was the preference for local (or national) references, considered more appropriate [186].

Studies have been carried out to compared WHO growth standards with local/national growth references, considering different geographic areas with genetic and environment diversity. These studies reveal that the local references are more faithfully than the WHO standards [188] [198].

The WHO standards do not consider ethnicity, socio-economic status, feeding mode, or other factors. They consider a healthy and "common" population and those not appropriated to individuals that belong to a risk population segment, such as the poor, belonging to an ethnic minority, unemployed parents, or those with some particular health issue with a high influence on the child [186] [189].

As mentioned above the current standards present some limitations. These limitations can be overcome using personalized medicine, a concept that it's becoming more popular every day.

Personalized medicine has achieved better results than a "standard" medicine, including in children, and with the evolution of technology, it is expected to bring true revolution on patient care on the near future. This new approach to health care can bring considerable advances in disease prediction, preventive medicine, diagnostics, and treatments [199] [200].

But what is personalized medicine? The US National Institutes of Health, for instance, describes it as, "the science of individualized prevention and therapy". It can also simply be described as a way of synthesizing an individual's health history, including information family history, clinical data, genetics, or the environment in a way to benefit individuals' health and wellbeing [200] [201].

This research aimed to contribute to improving growth assessments on children, creating a new tool that combines the concepts presented above, namely expected growth and personalized medicine on children. Having a more accurate target height will allow paediatricians, and other health professionals, to apply growth standards in a more accurate way and provide a better assessment on children's development.

Until now, the growth standards have been applied by professionals by considering the actual "state" (growth measure) for the child and the existing methods described in the following section which addresses the area related work, which offer a less accurate prediction. This work has created a new tool that offers better outcomes and can be easily adapted to specific populations, by considering their specificities.

Artificial Intelligence and machine learning are contributing, through many studies and projects all over the world to more personalized medicine, and thereby delivering better outcomes. We believe this study constitutes one more step towards the medicine of the future.

The following chapters will present details about the analyse of the existing methods through related work research, and about the research conducted with the application of the PhD thesis framework.

6.2 Related Work

As discussed in the previous chapter, the height assessment on a child assumes critical importance when improving child health and wellbeing.

Prediction for adult height can be used both for diagnosis and following treatments, providing a valuable tool in paediatrics. For instance, it can help to detect growth hormone deficiency with hGH and follow treatment performance. In diagnosis, the height of the child can be compared with the so-called "target height", by considering the parent's height and child's gender [190] .

Currently, there are two known methods for assessing the "target height".

Horace Grey has defined the following model to forecast the height (when adult) for a child [202]:

$$Female\ height\ (inches) = + \frac{father\ height\ \times {}^{12}/_{13} + mother\ height}{2} \quad (6.1)$$

$$Male\ height\ (inches) = + \frac{father\ height\ + mother\ height \times {}^{12}/_{13}}{2} \quad (6.2)$$

Other commonly used model is the Mid-Parental Method (described by Tanner) [203] [204]:

$$Female\ height\ (inches) = + \frac{father\ height\ \times + mother\ height}{2} - 2{,}5 \quad (6.3)$$

$$Male\ height\ (inches) = + \frac{father\ height\ \times + mother\ height}{2} + 2{,}5 \quad (6.4)$$

There is a lot of information regarding the applicability of the methods, but information about the studies' bases and characteristics are scarce, probably because they have been performed some decades ago. Still, it was possible to comprehend that both methods are based on statistics approaches where a simple regression method is applied to the population and the "optimal" formula is calculated. In a generic way, these approaches try to find the mathematical formula that best fits the general characteristics of the population.

Horace grey approach reports to a study published in 1948 and is based on 53 collected cases. This method introduces a different weight on the mother's or father's height when calculating the output prediction. It emphasizes that a mother and father's influence is different for different child's genders [202].

The method described by Tanner was introduced in 1970 and has been a standard procedure for assessing children's growth since. It was not possible to find a description of the dataset used in the study, but it was still possible to understand that this method approaches the height of parents on the same proportion, given the emphasis on children's gender, which is critical to defining height [205].

As we observed, the methods did not introduce personalization to the individual level, besides the influence of gender, which is one of the advantages that machine learning can introduce in these areas of study.

There are other studies besides parent's height which use bone age, actual age, or child's current height. All these studies were based on longitudinal population projects [206] [207].

These projects, also called cohort studies, involve observing and measuring population variables (the same individuals–cohort) for long periods, which can involve years or decades.

In the current study, we will only focus on information gathered at childbirth, and as far as we know, the two methods described before are the only existing "tools" for predicting a child's height when adults. They are quite popular, and it is easy to find applications of these methods in several child development areas.

It's important to highlight that we have also researched on applications of machine learning methods that address "target height" prediction. We didn't find studies addressing this subject.

Knowing that this subject was never approached on this perspective represents, in our view, an opportunity, and was one of the main reasons for proceeding with this work. This study develops a new method in forecasting children's target height using a completely different approach based on data science tools.

6.3 Business Understanding

The main objective of this initiative is to improve the field of growth assessment by developing a new technique to predict a child's adult height, also known as the target height. The model output will be personalized to the child (person-centric), and it is expected to provide a more accurate prediction when compared with the existing methods, already described in the previous chapters.

This work also included the search for the best “way” to compare the SotA methods with the one proposed, which usually leads to a metric adoption. There are several methods for measure/comparing models’ performance, such as: accuracy, confusion matrix, and several associate measures such as sensitivity, specificity, precision, recall or f-measure (very useful when facing classification problems with imbalanced classes), MSE—mean square error, ROC—Receiver Operating Characteristic Curve, AUC—area under the roc curve, RMSE—root mean squared error, MAE—mean absolute error, R^2—R Squared, Log-loss, Cohen’s kappa, Sum squared error, Silhouette coefficient, Euclidean distance, or Manhattan distance. [208] [209] [210] [211].

The various methods/measures present different approaches for different problems (regression, classification, and clustering) with advantages and drawbacks. Accuracy is a simple method (easy to compute and low complexity) able to discriminate and select the best (optimal) solution. It has constraints when dealing with classes, and when classes are not balanced may lead to sub-optimal solutions. Our initiative does not embody a classification problem, but rather a regression problem, and our goal is to provide better performance with our predictions. Given the characteristics of the challenge, it was decided that accuracy was the metric that best served the objective.

It was measured how well the related work "target height" prediction models performed, and the results are shown in Table 6-1. The Mid-Parental Method yields the best accuracy, with a 59.53% accuracy rate.

Table 6-1 Models found in the literature research (the challenge reference models)

Reference models	Mean Squared Error	Accuracy
Horace Grey	20.6983	58.55%
Mid Parental	5.2836	59.53%

In summary, the business goal was to provide a better method of estimating the “target height”, which leads to the need to outperform the SotA models, as thus improving the available tools in the growth assessment field. The accuracy was adopted as the evaluation's metric.

6.4 Data Understanding

For the proposed study, was decided to use a dataset based on the famous study of Francis Galton (an English statistician pioneer, who developed concepts such as correlation, quartile, percentile, and regression) that was used to analyse the relationships between child’s adult heights and their parents' adult heights [212].

The dataset is based on the observation of 928 children and the corresponding parents (205 pairs), where several variables have been recorded, such as father and mother height or child gender and adult height [213].

Galton's study introduced the idea of regression analysis and contributed significantly to the understanding of multi-variable relationships. The study unveiled that parents' heights were distributed according to two Gaussian distributions (also known as normal distribution), one for the mother and one for the father, which, when combined, resulted in a bivariate normal distribution. It also has proved that adult children's height could be described by a Gaussian distribution [214].

The dataset used to this experiment is composed of 898 cases and the measuring unit is the inch. The dataset file includes the following information:

- Family: Family identification, one for each family.
- Father: Father's height (in inches).
- Mother: Mother's height (in inches)
- Sex: Child's gender: F or M.
- Height: the child's height as an adult (in inches).
- Nkids: the number of adult children in the family, or, at least, the number whose heights Galton recorded.

Since the dataset's original measurement unit was the inch, it was decided to conduct the research using that measurement unit to prevent potential noise from rounding the numbers. One inch is equal to 2.54 centimetres.

Figure 6.1 plots the distribution of heights for mothers, fathers, child's (regardless of gender), female child's and male child's. The horizontal axis presents the height, measured in inches, while the vertical axis shows the distribution density or put another way, the probability of a given value on the horizontal axis. The following conclusions can be drawn from analysing the graphic:

- Child's gender it's a feature with high impact on height forecast.
- All "segments" heights (mothers, fathers, child's, female child's, and male child's) correspond to Gaussian distributions.
- Female child's and mother's height distributions are similar. The heights of the male child and parents also match.

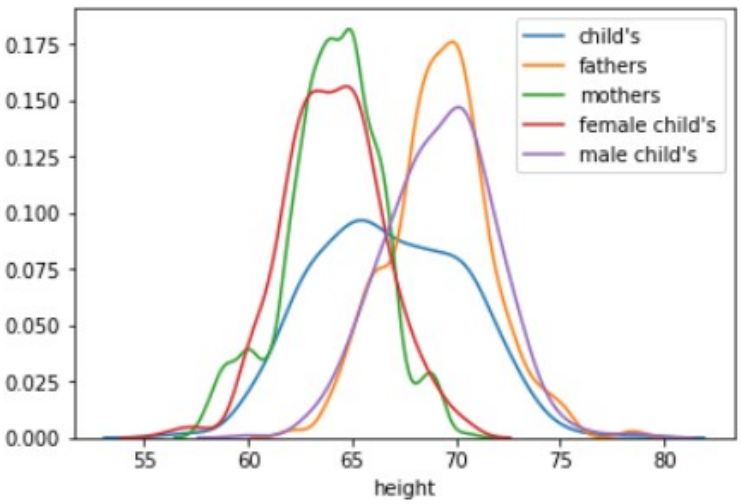


Figure 6.1 Distribution of heights, including for mothers, fathers, child's, and both male and female child's

The following graphs make it possible to visualize how the heights of the fathers, shown on the horizontal axis, and the child heights, shown on the vertical axis, relate to one another. In the first chart, Figure 6.2, all dataset was included, and two nuclei can be observed. It can be quickly inferred that these nuclei result of two "events" existing in the relationship between father's and child's heights.

To better understand these two different densities, we decided to split the dataset into male and female child's, which can be observed in Figure 6.3 and Figure 6.4, respectively.

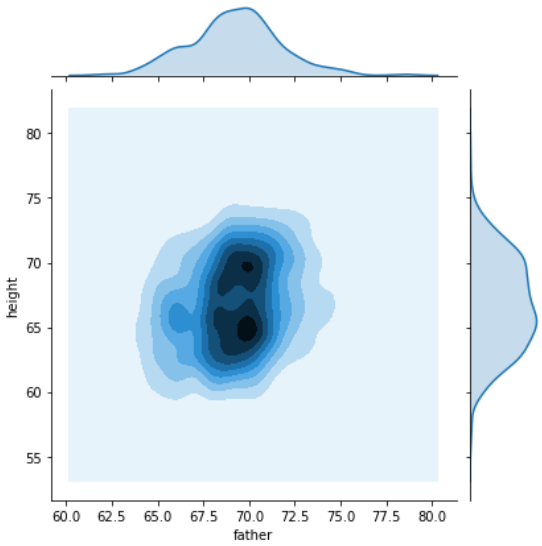


Figure 6.2 Relation between father height and child height (regardless of gender).

Figure 6.3 and Figure 6.4 clearly identity a relation of father's height with child's height that is different when considering the child's gender. A child will be taller when male and shorter when female for the same father height.

The same exercise was performed using the mother's height (rather than the father's) and similar shapes were obtained, but different influence dimensions were revealed. This analysis shows a high degree of influence caused by the child's gender as well as different influences from the mother and father.

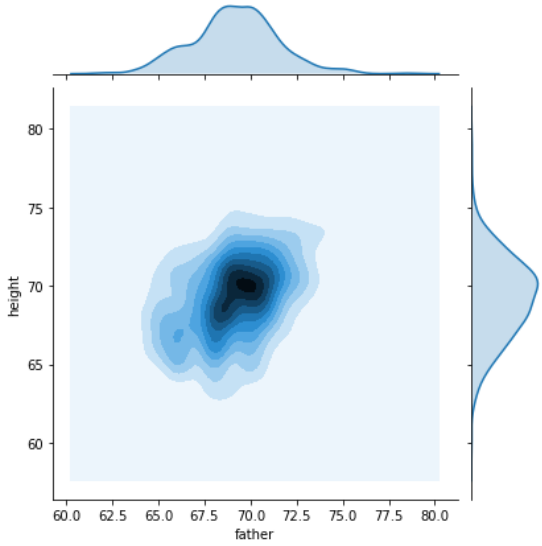


Figure 6.3 Relation between father height and male child height.

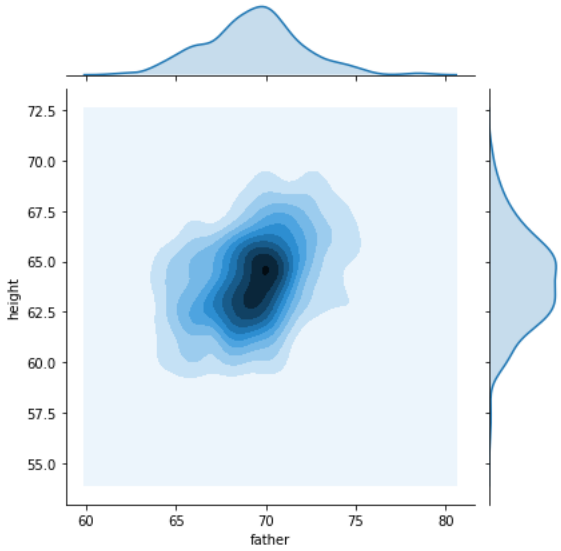


Figure 6.4 Relation between father height and female child height.

For further understand this relationship it was decided to conduct a correlation analysis to understand how mother and father influenced the child's height variations. For the purpose, the Pearson correlation measure was adopted.

It was also decided to proceed this analysis dividing the dataset in several variants, namely analysing children's height by gender, considering the corresponds mother and father separately, and according to parents to height distribution quarters. The results can be found in Table 6-2.

Table 6-2 Height's correlation considering different perspectives, namely: child's considering gender and regardless of gender; father's and mother's height quarters

Parent's	Dataset split	Child's	Male	Female
Father	All	0.275	0.391	0.459
	quarter 1	0.115	0.055	0.130
	≤mean	0.215	0.263	0.216
	>mean	0.144	0.255	0.356
	quarter 4	0.044	0.156	0.187
Mother	All	0.202	0.334	0.314
	quarter 1	0.180	0.275	0.236
	≤mean	0.134	0.197	0.291
	>mean	0.217	0.249	0.351
	quarter 4	0.159	0.311	0.247

The following conclusions were drawn based on the results:

- The relation between a parent's height and a child's height varies depending on the child's gender. This is evident for instance when considering (all) father influence on female (0.459) and male (0.391) children. We observe a stronger effect on females.
- Father's height and mother's height have different weights/influences on the child's height. This can be observed, for example, in female's child's where the correlation with the father height is 0.495, but only 0.314 for the mother.
- The parent's height influence on child's height it's not constant, it varies among the height ranges / splits.

Still in this data understanding phase context, it was considered relevant to perform a linear regression analysis. Again, some variations were taken into account, including considering for the horizontal axis the sum of parent's heights (Figure 6.5), the mother's height or father's height individually; and considering for the vertical axis the child's height (regardless of gender), or considering the child's gender (Figure 6.6 for female gender).

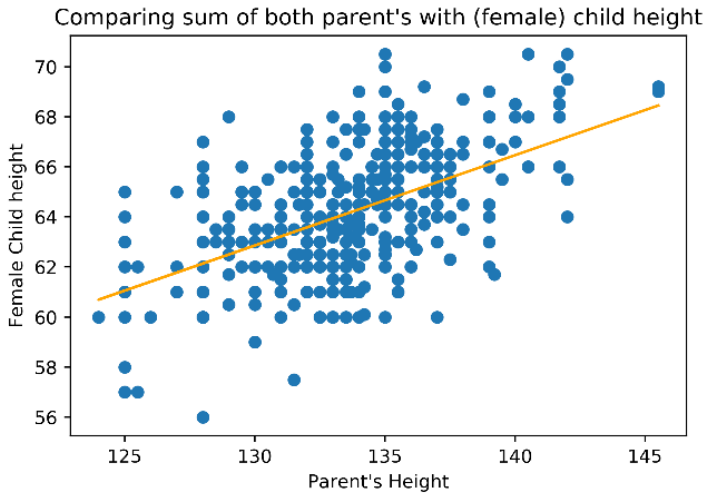


Figure 6.5 Regression analysis considering (both) parents and female child's height

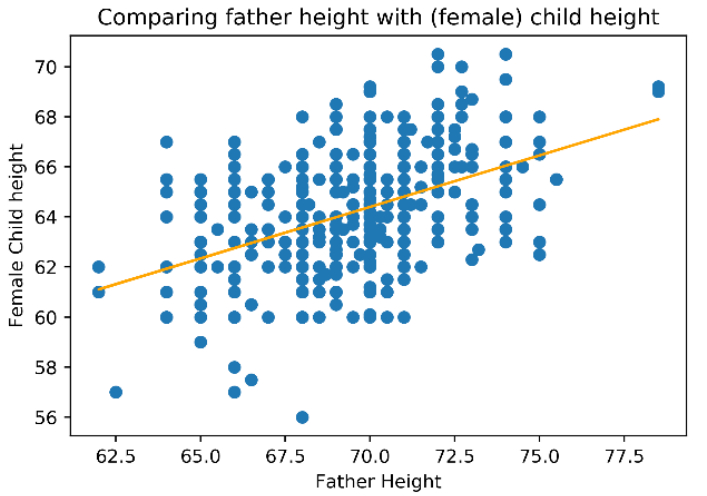


Figure 6.6 Regression analysis based on height of the female child's and the father's

We found a multilinear regression, where the child's height depended on the individual's variable mother and father heights. We also noticed a high variance in the relationships as it could easily be observed on the plots, meaning same parent's heigh can correspond to a significant child's height range.

Verifying the dataset for errors, duplicates, missing or incorrect data, and other data issues is another step included in this phase. In Table 6-3, it is possible to observe one of the tools used for verifying data, such as the number of rows, the fact that each row contains values for each feature, or number of unique values for each feature.

Table 6-3 Dataset resume

Metric	Family	Father	Mother	Sex	Height	Nkids
count	898	898	898	898	898	898
unique	197	-	-	2	-	-
top	185	-	-	M	-	-
Freq	15	-	-	465	-	-
Mean	-	69.23	64.08	-	66.76	6.13
Std	-	2.47	2.30	-	3.58	2.68
Min	-	62.00	58.00	-	56.00	1.00
25%	-	68.00	63.00	-	64.00	4.00
50%	-	69.00	64.00	-	66.50	6.00
75%	-	71.00	65.50	-	69.70	8.00
max	-	78.50	70.50	-	79.00	15.00

There was no need for data cleaning or other similar tasks on the dataset. This was anticipated given that other researchers had previously used our dataset for other scientific purposes.

6.5 Data Preparation

Feature selection can be performed using several techniques. These techniques can be classified, in a first view, into supervised, unsupervised, and semi-supervised approaches. Inside the supervised segment, the category applied to this particular challenge, we can find different categories such as filter methods, wrapper methods or embedded methods. Information gain, Relief, Fisher score, gain ratio, and Gini index are a few examples of feature selection techniques that fall into these categories [215] [216] [217].

The "family" feature describes the family to which the child belongs, in accordance with the previous data understanding phase. The machine learning models can use, during the training process, the “family” feature to extract the brothers and sister heights when adults and use this information to forecast the child’s height. In the current scenario, was decided not to use brothers’ and sisters’ heights since, in real cases, they are unlikely to be adults yet, and consequently the brother’s height will not be available at that forecast moment. The family column was thus left out of the dataset to be used for training, and as a result, family relations are not considered.

The nkids feature, which refers to the number of brothers of an individual, has also been excluded since machine learning can make use of this data in some way. This information is invalid because, for instance, it is unknown at the time of the first brother's birth how many brothers he will have in the future, although the information is stated in the dataset.

The other features were likewise examined in the scope of the feature selection tasks. It is important to recall that the previous data understanding phase concluded that each of the other features appeared to present, in some degree, influence in the prediction task goal. Furthermore, the algorithms used on the study wrap feature selections techniques and those can address finer feature selection automatically. Considering these facts, it was decided to include all the remaining features in the modelling phase.

Many machine learning algorithms struggle (or are even unable) to handle categorical variables. In order to apply a wider variety of algorithms, the gender feature was transformed into a binary feature defining whether the child "is male?".

In the process of supervised machine learning is necessary to divide the entire dataset on training and test sets. Frequently, an 80/20, 90/10, or 70/30 ratio is used. In this case was decided to use the first option. Once divided it’s crucial to ensure that both (sub) datasets maintain the original dataset shape This was assured as it can be observed in Table 6-4 and Table 6-5.

Table 6-4 Model's training dataset

Metric	Father	Mother	Sex	Height
count	718	718	718	718
unique	-	-	2	-
top	-	-	M	-
Freq	-	-	370	-
Mean	69.36	64.11	-	66.84
Std	2.51	2.29	-	3.57
Min	62.00	58.00	-	56.00
25%	68.00	63.00	-	64.00
50%	69.50	64.00	-	67.00
75%	71.00	65.50	-	70.00
max	78.50	70.50	-	79.00

Table 6-5 Testing dataset for model's performance evaluation

Metric	Father	Mother	Sex	Height
count	180	180	180	180
unique	-	-	2	-
top	-	-	M	-
Freq	-	-	95	-
Mean	68.73	63.99	-	66.44
Std	2.20	2.38	-	3.62
Min	62.50	58.00	-	57.00
25%	67.37	63.00	-	63.50
50%	69.00	64.00	-	66.25
75%	70.00	65.50	-	69.05
max	75.00	70.50	-	76.50

6.6 Modelling

The modelling phase starts with the identification of the algorithms to apply to the specific problem. A common strategy was used for the current challenge, namely the application of strategies that have shown good results in other challenges of a similar nature. After doing a survey of the literature, the following "popular" machine learning algorithms were selected [218] [219].:

- Linear regression;
- SVR—Support Vector Regression (several kernels);
- Lasso;
- Elastic Net;
- Ridge Regression;
- K-Neighbours Regression;
- Bayesian Regression;
- AdaBoost Regression;
- SGD—Stochastic Gradient Descent Regression;
- Decision Tree Regression;
- XGB—Extreme Gradient Boosting Regression and;
- LGBM—Light Gradient Boosting Machine Regression

After the algorithm's selection, the training phase was performed. For this task the training dataset, described in the previous section, was submitted to every of the identified algorithms. The setup of development environment for the training process was: Jupyter notebooks supported in Python 3.6 and the Python libraries Sklearn, XGBoost and Lightgbm.

Regarding the algorithm's "configuration", we have chosen to begin our modelling phase using the default settings (hyperparameters) set by the aforementioned libraries. Table 6-1 provides an overview of the training process outcomes.

Table 6-6 Machine learning algorithms performance with default hyperparameters

Algorithms	Mean Squared Error	Accuracy
Linear	5.1781	60.33%
SVR RBF	6.1269	53.08%
SVR Linear	5.2167	60.04%
SVR Poly	Invalid output	----
SVR Sigmoid	13.262	1.58%
Lasso	9.8600	24.47%
Elastic Net	9.9896	23.48%
Ridge	5.1829	60.30%
KNeighbours	6.3785	51.14%
Bayesian	5.1805	60.32%
AdaBoost	5.3941	58.68%
SGD	Invalid output	----
Decision Tree	5.5017	57.86%
XGB	4.8290	63.01%
LightGBM	4.7938	63.28%

Using this approach several algorithms (such as SVR Linear, Ridge, Bayesian, or Linear regression) were able to overcome the accuracy of the existing models The majority of these algorithms only saw modest improvements, but XGB and LightGBM stood out with a considerable improvement.

SVR Poly and SGD algorithms were unable to converge using the current strategy (default parameters and no data input custom transformations for modelling) and were unable to produce a valid result.

The study's goals, which included creating a model to forecast target heigh and improve the performance provided by the existing models, namely Horace Grey and Mid Parental method, were achieved. Nevertheless, the contribution of this study also entails providing a tool that can advance the growth assessment field, and thus it was important to keep working for the highest level of accuracy.

The use of (input data) scaling and hyperparameters tuning are two of the most popular methods for enhancing the results of machine learning training.

Multiple scaling techniques, including StandardScaler, MinMaxScaler, MaxAbsScaler, RobustScaler, and (simple) Scale, have been tested during the modeling phase. The MaxAbsScaler, which rescales data so that all feature values are in the compromised range between [0, 1], produced the best results.

Table 6-7 Machine learning algorithms performance after applying MaxAbsScaler scaling technique

Algorithm	Mean Squared Error	Accuracy	Comments
Linear	5.1781	60.33%	No change
SVR RBF	5.3131	59.30%	Improved (±6%)
SVR Linear	5.2398	59.86%	Deteriorate
SVR Poly	5.7943	55.61%	Valid output
SVR Sigmoid	5.4715	58,09%	Improved (±56%)
Lasso	10.6798	18.19%	Deteriorate
Elastic Net	10.8461	16.92%	Deteriorate
Ridge	5.1916	60.23%	Deteriorate
KNeighbors	5.4779	58.03%	Improved (±7%)
Bayesian	5.1798	60.32%	No change
AdaBoost	5.2607	59.70%	Improved (±1%)
SGD	17.3546	32.93%	Valid output
Decision Tree	5.5299	57.64%	No change
XGB	4.8290	63.01%	No change
LightGBM	4.7938	63.28%	No change

Although the application of this method increased some algorithms' accuracy, the height prediction forecast in general, which is the goal of the study, have not been improved.

It is relevant to highlight that the technique was able to solve the previously verified convergence issue with SVR Poly and SGD, enabling them to produce a valid output.

6.7 Architecture Optimization

The results achieved by the modelling phase were already interesting and the primary initiative goals had already been met, but as previously discussed in this thesis, the architecture optimization offers the opportunity to improve model performance, and consequently the overall results.

For the algorithms previous discussed the optimization consists in the hyperparameters tuning and in this challenge two strategies were considered: Grid search and Random search. A Grid search approach requires a lot of resources, including both time and energy to compute all possible hyperparameters combinations, for all considered algorithms. Considering the initiative resources constraints, a random search strategy was considered for the optimization phase, with 100 iterations per algorithm.

The outcomes of the optimization are presented in Table 6-8.

Table 6-8 Machine learning algorithms performance after hyperparameters tunning

Algorithm	Mean Squared Error	Accuracy	Comments
Linear	5.1781	60.33%	No change
SVR RBF	5.3131	59.30%	No change
SVR Linear	5.2021	60.15%	Improved (±0.5%)
SVR Poly	5.2965	59.42%	Improved (±4%)
SVR Sigmoid	5.3044	59.36%	Improved (±1%)
Lasso	5.1697	60.40%	Improved (±42%)
Elastic Net	5.1697	60.39%	Improved (±43%)
Ridge	5.1782	60.33%	Improved (±0.1%)
KNeighbors	5.2945	59.44%	Improved (±1%)
Bayesian	5.1798	60.32%	No change
AdaBoost	5.2607	59.70%	No change
SGD	5.1871	60.27%	Improved (±28%)
Decision Tree	5.5339	57.60%	No change
XGB	4.7125	63.90%	Improved (±1%)
LightGBM	4.6159	64.64%	Improved (±1%)

The accuracy of many algorithms has increased thanks to hyperparameter tuning (Figure 6.7), but most importantly, the study's primary objective—improving children's "target" height prediction accuracy—has been met.

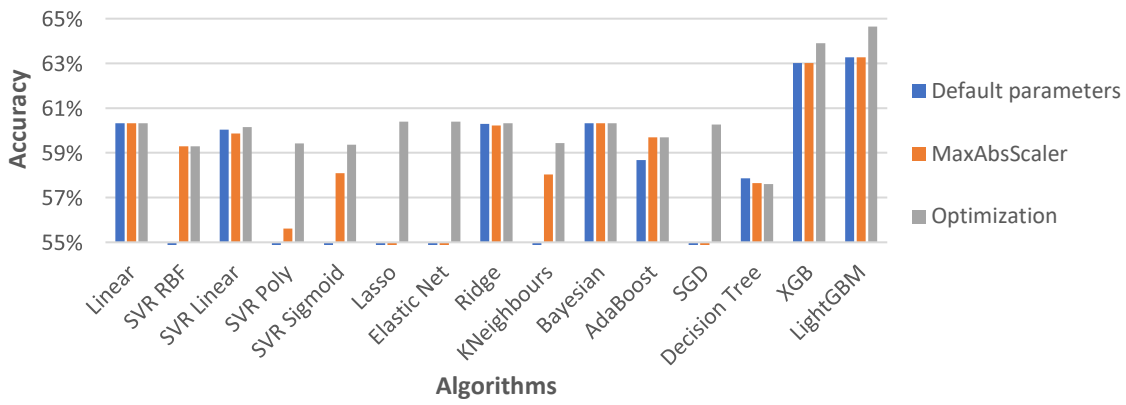


Figure 6.7 Accuracy improvement in the modelling and optimization phases

In addition to the performance improvement output, it is also interesting to analyse the costs associated with this task. As a result, keeping track of the computing resources needed for each iteration of the algorithms were part of the task. Due to the variable nature of CPU usage, it was decided to analyse the CPU time mean and standard deviation while considering all iterations.

The task was carried out by a machine with the following features:

- Processor: Intel Core i7-3770 3.40GHz;
- RAM: 8 GB;
- Operating System: Windows 10, 64 bits, x64-based processor.

Table 6-9 Resources consumption for the machine learning algorithms optimization

Algorithm	CPU Time Mean	CPU Time Std	Memory Consumption (MB)
Linear	1.87 ms	20 µs	138.44
SVR RBF	197.00 ms	3.33 ms	141.62
SVR Linear	101.00 ms	1.66 ms	140.76
SVR Poly	>10.00 m *	*	*
SVR Sigmoid	11.60 ms	211.0 µs	156.61
Lasso	1.50 ms	26.2 µs	156.87
Elastic Net	1.65 ms	58.1 µs	154.52
Ridge	1.18 ms	39.8 µs	154.91
KNeighbours	868 µs	34.8 µs	156.00
Bayesian	1.44 ms	33.7 µs	156.02
AdaBoost	15.00 ms	260.0 µs	156.19
SGD	5.85 ms	68.4 µs	156.32
Decision Tree	1.36 ms	28.8 µs	156.57
XGB	26.6 ms	1.01 ms	157.84
LightGBM	107.00 ms	1.26 ms	163.16

* the purpose of this analysis was to compare execution consumptions which are typically measured in milliseconds scale. Support Vector Regression (SVR) Poly algorithm took longer than 10 minutes to process, and thus and we have simply decided to label it as “long time processing” (>10 m).

6.8 Evaluation

The modelling phase was responsible for developing a few models able to forecast the target height. The following phase of optimization was responsible for improving the model’s performances. Considering the results obtained, it is possible to state that the machine learning goal was attained with success.

Ten of those models have outperformed the results expressed by the related work reference model, Mid-Parental Method, and two of the produced models, XGB and LightGBM, stood out for having significantly improved the performance. The second objective outlined in the business understanding phase was accomplished by these two models, increasing the (until now) existing prediction accuracy.

It is helpful to look at the results of the approaches to understand why the best model, LightGBM, was able to surpass the accuracy of the Mid-Parental method (the best model on related work).

In the first graph, Figure 6.8, the results for both methods are compared (vertical axis) while considering the sum of the heights of the parents (horizontal axis). In the Figure 6.8 we can find both outcomes for the model LightGBM (red) and Mid-Parental method (green). Real values for the child's height were added to the Figure 6.9 (in blue).

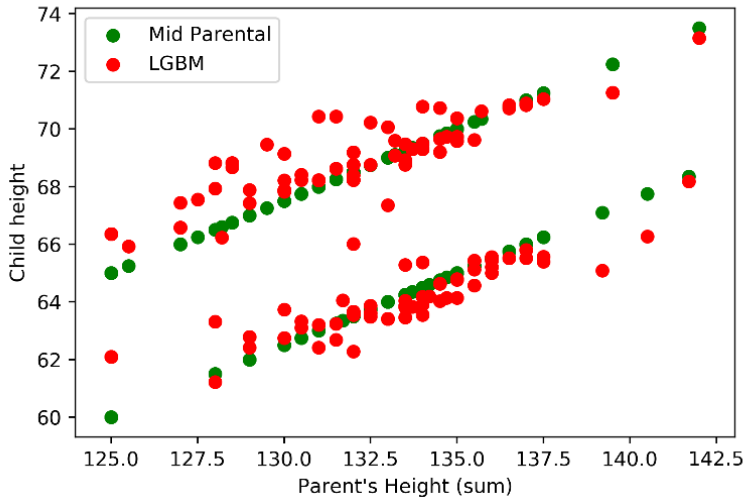


Figure 6.8 LightGBM (red) and Mid-Parental (green) prediction outputs.

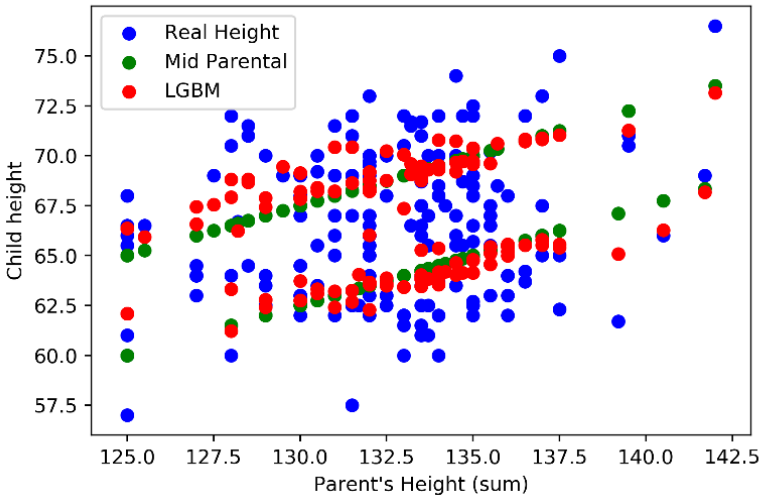


Figure 6.9 LightGBM (red), Mid-Parental (green) model's prediction and the child's actual height (blue)

As shown in the preceding figures (Figure 6.8 and 6.9), it is possible to verify that the Mid-Parental always predicts the same child height, for the same gender, when using the same parent's height sum. The LightGBM model presents different values for the same circumstances. Comparing the outputs of the two models, LightGBM and Mid-Parental, with the actual values makes it is easy to understand that the first model is able to produce predictions that are more closely related to the real values (LightGBM MSE 4.6159 Vs Mid Parental 5.2836).

The following section, 6.9 Explanation, provides a more detailed analysis of the assumptions underlying the outcomes generations.

After the results analysis, every step of the process was reviewed to make sure there were no potential mistakes, leaks, or other failures. There were no issues found, so it was decided to move on to the next phase.

Before advancing to the new phase and considering that in this case resources consumption for optimization were also considered, the results should be also evaluated.

Regarding the memory consumption we can observe a low variance (18%) between the several algorithms, with a minimum use of 138 Mb and a maximum of 163 Mb. In our option, this is not a significant difference, and even if the processing required larger files (for instance for other kind of challenges), the resources currently available could process all the data in a short period of time.

The CPU time for running the optimization can be split into four classes. The algorithms on the milliseconds range (e.g.: Linear, SGD or KNeighbours), those on the tens of milliseconds range (e.g.: AdaBoost or XGB), the class for the hundreds of milliseconds (e.g.: SVR RBF or LightGBM), and finally the SVR Poly on the long-time processing class.

The CPU consumption is directly proportional to the algorithm's complexity (number of operations). It was expected to have a low processing consumption for a "simple" linear, whereas support vector regressions were anticipated to require more processing power. It is important to keep in mind that linear regression is a simpler method that builds a linear equation (e.g.: $y = \beta 0 + \beta 1 \times 1$) that best fits the linear relationship between a set of input data to predict the output value, in contrast to SVR algorithms which attempt to fit data using a hyper-plane where the support vectors can take various shapes (including nonlinear). These shapes can vary according to the kernel defined as parameter for the SRV [220] [221].

Focusing on the algorithms XGB and LightGBM, that show the greatest improvements, we can see a significant difference in CPU consumption, with the second one taking four times as long as the first one.

Both algorithms use ensemble techniques based on gradient-boosted trees. In contrast to LightGBM, which grows trees vertically, XGB grows trees horizontally like most tree-based algorithms. This means that, while the other takes into account the level (level-wise), LightGBM grows from the leaf with the greatest loss (leaf-wise), which can reduce loss and lead to better accuracy, as it is demonstrated by the achieved performances [222] [223] [224].

LightGBM was developed after XGB and uses a novel sampling method called Gradient-based One-Side Sampling. It was developed with the purposes of achieving faster training, lower memory consumptions and better accuracy. Our findings support the accuracy hurdled, but we see the opposite for resource consumption. That is easily accounted by the fact that LightGBM was designed for large datasets, which was not the case in our study.

6.9 Explanation

The current phase is responsible for "diving" into the model and understand how the outcomes are generated. The most performant model it is not interpretable by design, meaning it is not possible to directly extracts information that explains the produced forecasts. Thus, the use of explanation techniques was required. In this instance, SHAP methods offered the best choice.

The gender feature is the focus of the first analysis. Figure 6.10 depicts how gender affects the LightGBM model. Male children are represented by the colour cadet blue (feature is male = true), while female children are represented by the colour dark orchid (feature is male = false). Although the influence of gender is not constant at five inches, it is generally centred at 4.9 inches with some minor variation. The variation is 0.5 inches on the female (dark orchid area) and 0.7 inches on the male (cadet blue area. No trend could be inferred from the variation.

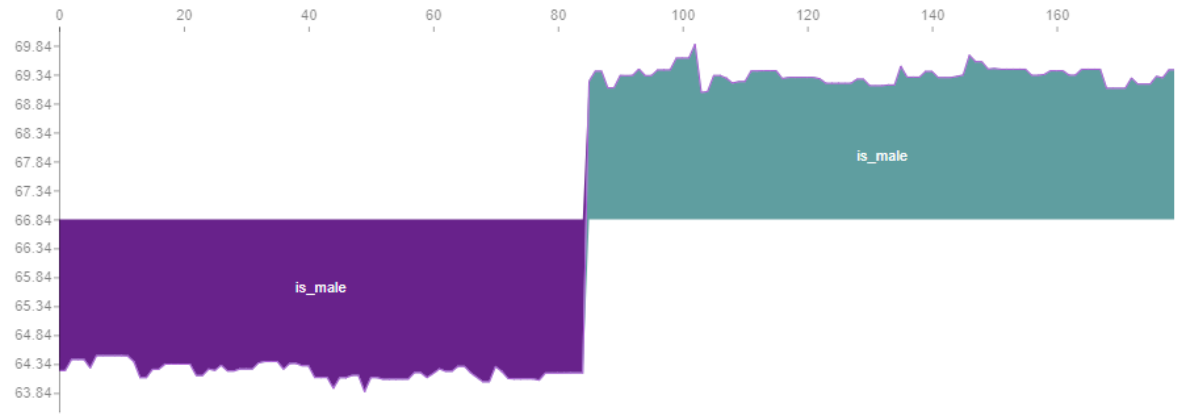


Figure 6.10 Effect of gender on predicting a child's height (dark orchid for the female gender and cadet blue for the male gender)

Previously, in the data understating phase, it was observed that the influence of mothers and fathers varies depending on the child's gender, and for different mother's and father's heights ranges. In Figure 6.11, it is possible to confirm that the produced model can successfully capture and embed that information.

The model's explanation demonstrates how the influence of mothers' and fathers' heights varies depending on the gender of the child. It is confirmed that fathers, on average, have a greater influence on a child's target height, with a focus on female children since its influence is more pronounced. Mothers have less impact on both genders, and contrary to the father's, the gender that is most impacted is the male gender.

These gleaned insights support the data exploration's findings, which revealed, for instance, that the correlation between the father's height and the height of the female child was 0.459, while the height of the male child correlated at 0.391.

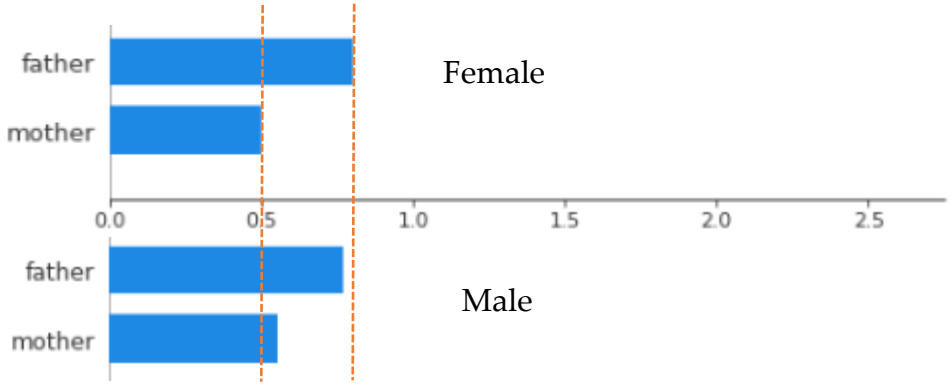


Figure 6.11 Parent's influence on child´s height, considering the child gender

While in Figure 6.11 it is possible to observe the mean influence of parent's height, Figure 6.12 provides the specific influence of mother and father heights for the female child's, which delivers a different and more in-depth viewpoint.

The records for every female child are shown in the graph below, from tallest to shortest. The red colour indicates a favourable influence on the child's height, while the blue colour indicates a negative influence.

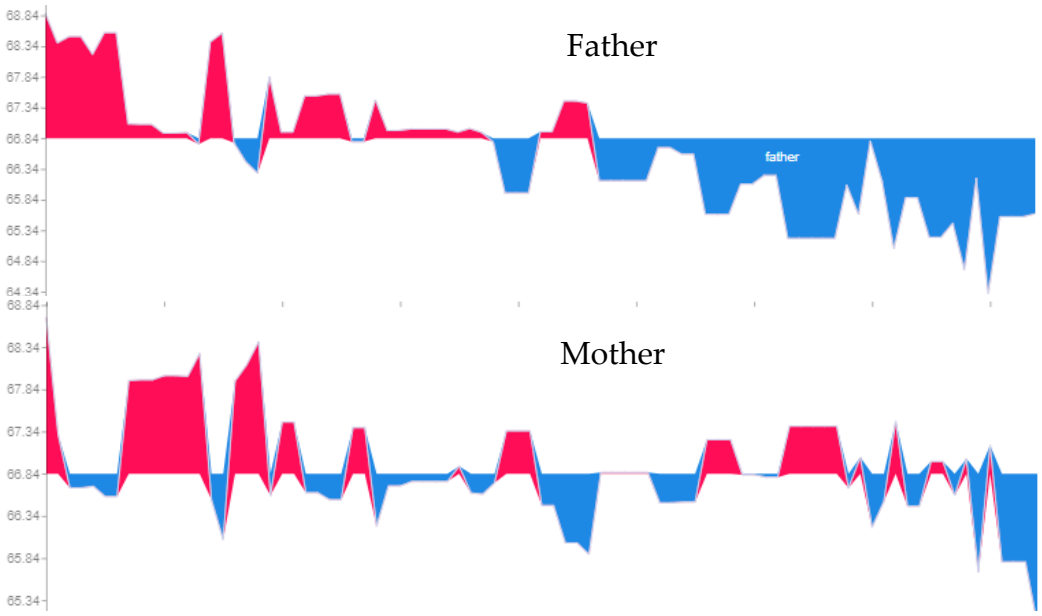


Figure 6.12 Influence of parent's height on each female child's height (red for positive influence and blue for negative influence).

The most significant finding from the previous graph is that parents' height has a very different impact on their female children's height. The father's height mainly has a positive influence on the tallest female child's height (above the mean), whereas mothers' height does not have the same effect. Without a clear trend, the mother's influence exhibits a lot of fluctuations.

This variation was already apparent during the data understanding stage, where there were flotations in the correlation between mothers and female children.

We can conclude from the figure above (Figure 6.12) that the model can account for the various influences of the parents' heights in each of the cases. This also explains how the model was able to enhance the performance of the height target forecast.

6.10 Deployment

After achieving the objectives set forth for this study, we believe it was important to provide a simple means by which clinical professionals, particularly those who concentrate on growth assessment, can benefit from this evolution. To do this, we have developed a simple interface (Figure 6.13), linked to the model by a web service, that enables quick introduction of the inputs and generation of the output prediction. This interface, model, and service can be incorporated on an already-existing IT infrastructure in healthcare facilities like hospitals and clinics.

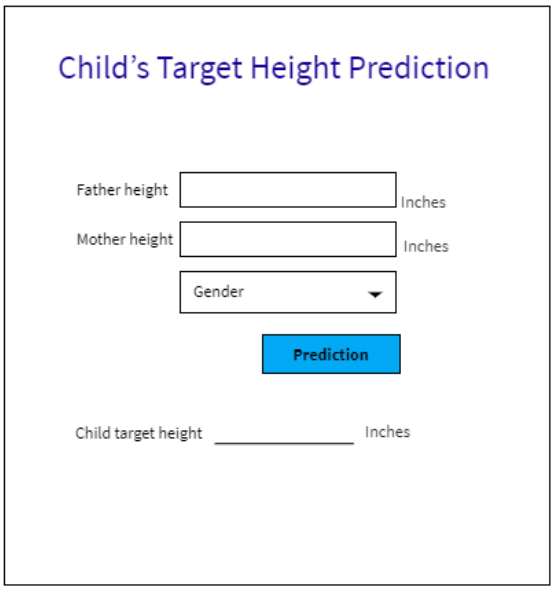


Figure 6.13 Interface for Child's Target Height Prediction tool

We hope that this solution can become an interesting choice in growth assessment.

6.11 Research Results

This study confirms many of the facts that were previously known through SotA research and other studies on children growth area. Specifically, the effect of parental height on the height that children will attain, as well as a difference in growth expectations based on the child's gender, with a male child clearly tending to be taller!

The innovation introduced by this study, for the field under investigation, was the way in which the target height was specifically influenced by an individual's characteristics, including hereditary factors like the height of the mother and father. Now, contributions of the parents' height and gender features are more clear / accurate.

It is now understood how gender affects target height (4.9 inches difference between males and females) as well as the variance between individuals of the same gender: 0.7 inches for male child's and 0.5 for female.

It is also known in which dimension father and mothers influence the child height. In contrast to the SotA models, which have different perspectives as discussed in the related work section, the research makes clear that the father's height always has a greater impact than the mother's height. The model forecast should be used to obtain the most accurate forecast because the relationship between parental height and the corresponding height for children is not linear.

Although the study features largely explain the child's target height, child's environment features are also relevant, and that is what explains most of the variability not captured by the model. With the inclusion of more variables, the model research could become (even) more precise in the future.

The model also has the advantage of being adaptable to any specific population; this only requires retraining the model with data from the target population. Or, from a different angle, it demonstrates the potential for creating initiatives in data science related to health and proved the ability to discover knowledge in datasets that already exist.

Although the dataset in this instance was relatively simple, the method can undoubtedly be used to solve other problems, including those involving larger and more complicated datasets.

6.12 Conclusions

To improve the growth assessment field, the research evaluated the capacity to predict children's adult height. The challenge was to use machine learning methods to overcome the existing predictive models.

Following the results, the research was able to find better approaches to the children's target height forecast. Most of the tested approaches saw only modest improvements, but using the XGB and LightGBM algorithms led to significant improvements when compared to the (up until now) existing models.

By analysing the developed LightGBM model, we understood that it was able to incorporate the varying degrees of influence that fathers' and mothers' heights had on their children, as well as its effects on different genders. This was a major contribution for the output forecast and a novelty when compared to the earlier models.

As a step forward toward a more personalised medicine, the study created a straightforward "tool" to assist paediatricians and other health professional interested in the growth assessment area. Furthermore, the tool can be easily customized to any specific population.

Comparing the final model to real values still reveals some variation. This variation can be reduced by adding additional significant variables to the model, as the height of the grandparents, existing pathologies, socioeconomic status, and alimentation access/quality. These are great opportunities for future work in the area.

In the context of this PhD thesis, expanding this initiative with a more exhaustive dataset would also be a fantastic opportunity. A collaboration offer was made to the Portuguese government services, namely Instituto dos Registos e do Notariado (IRN), and a partnership was achieved. Currently, we are still waiting for the dataset that incorporates the Portuguese population information.

CHAPTER 7 - Applied Data Science: Length of Stay in Neonatal Intensive Care Units

This use case presents a length of stay (LOS) analysis on neonatal intensive care units, extracting hidden knowledge from a 10 years dataset of all Portuguese neonatal intensive care units. In contrast to conventional studies that rely on statistical analysis, this investigation uses data science techniques, including machine learning hyperparameters optimization automatization and eXplainable Artificial Intelligence tools, to understand the significance of factors and in which way they affect LOS.

The initiative presents a list of factors that contributes to LOS in Portuguese neonatal intensive care units which are analysed and ranked by relevance and impact.

The analyse is performed according to the already defined methodology (3.2 Thesis Framework), using a combination of data science techniques that can also be applied to other clinical domains. The work is pertinent for both offering insightful information for enhancing investment in LOS reduction (one driver for the initiative), as well as for providing methods for handling data in other health-related challenges.

Part of this research was already published in an IEEE conference, under the title “Length of stay analysis at neonatal care units with data science-preliminary results”. The research's final findings will be included in a second publication that is currently being prepared [2].

7.1 Introduction

Length of stay often just defined as LOS in the health area, is the concept that describes the time spent in the hospital receiving medical care and is defined as the total days of hospitalization per patient from admission to discharge [225] [226] [227].

The need to stay in a hospital for medical care is usually the result of an unhealthy situation and/or a context of adverse factors. As a result, being hospitalized is an undesirable situation for all parties involved, from patient to medical staff, or the care payers – insurers, government services, or others. It is a common interest to all parties to reduce the LOS, regardless of age or medical condition, although there are cases in which hospitalization can be maintained due to prevention.

LOS prediction in general is a very interesting area as it can help health professionals to monitor and structure health care services and manage resources. It is also a useful tool for health consumers (e.g.: patients and/or family members) since it can help manage expectations which contributes to reducing anxiety in a quite stressful period [227] [228].

LOS factors and their impacts can vary widely according to the medical situation and patient.

The first step in addressing a LOS "problem," such as the current use case of LOS analysis in Neonatal Intensive Care Units (NICUs) that may contribute to efficient changes, is to collect data that can have an influence on the situation. Usually, this type of information is acquired by clinicians by observation and/or with the collaboration of statistics professionals who develop statistical analysis resorting to several statistical tools and methods [228] [229] [230].

The present initiative uses the first ten years of records of the National Registry of Infection in Portuguese NICUs which is composed by over 50k records, among all Portuguese NIUCs.

This investigation work includes a predictive approach to the neonatal LOS in NICU, which makes use of machine learning techniques. However, the focus of the study is not the model accuracy or capacity “as usual” (although of course it is very important as a base to extract insights) but otherwise how the newborns’ characteristics and involved factors contribute and impact the LOS. This is a distinguishing trait of this study, as other related LOS studies that use machine learning (ML) focus solely on the model and the model accuracy.

This analysis aims to be a trigger to decision-makers choose alternatives for the investment in LOS prevention, contributing to newborns health and wellbeing, and consequently to a healthier society. At the same time, it can contribute for optimizing the use of scarce medical resources [227] [231] [232].

The present work involves several data science techniques, such as improving the model's performance using machine learning hyperparameters optimization automatization (both structural/architecture and training hyperparameters), including Neural Architecture Search (NAS); and extracting more complete LOS causes insights, which includes knowing factors importance and impacts using eXplainable Artificial Intelligence techniques (XAI). The results are analysed, discussed and documented with a clinical team in a way to extract the best medical interpretations for the “technological achievements” [233] [234].

The study was approved by the Ethics Committee of the ISCTE - University Institute of Lisbon, with the number 60/2019, on the 11th of September 2019.

7.2 Related Work

Studies addressing LOS in neonates’, and in LOS in general, are frequently based on “traditional” statistical analysis where the output generally consists of univariable or/and multivariable analysis, multiple linear regression analysis, ratios, ranges, averages, correlations, among other statistical measures. One of the most used measures is the mean stay, which, by definition, is the mean number of days of hospitalization per patient, calculated by adding the days of hospitalization of each patient and dividing by the number of patients [225] [228] [229].

The clinician's experience and studies, including the papers approached in this introduction, were already able to verify that lower birth weight (BW) and gestational age (GA) have a strong inverse correlation with LOS, where a more immature newborn usually requires more clinical attention/care. This is also confirmed by the observed longer length of stays (LOS) due to the increased survival of very-low birth weight infants (VLBW<1500g) [229] [235] [236].

LOS, among other causes, can also be impacted by healthcare-associated infections (HAI), a major concern among clinicians, which beyond morbidity and mortality represent a waste of economic resources that may be of great impact. HAI is defined by the World Health Organization (WHO) as "an infection occurring in a patient during the process of care in a hospital or other healthcare facility which was not present or incubating at the time of admission [237] [238].

HAI it's also defined by NEO-KISS, the German registry of low birth weight newborn infants, as any infection starting 72h after admission that was not in incubation before admission. This definition was adopted in the Portuguese Registry [239].

Studies addressing LOS are frequently based on a quite reduced number of cases, concentrated on few NICUs (often a single unit). The studies present these populations as a limitation that can influence results. For instance, the research by M. Berry et al. "Predictors of mortality and length of stay for neonates admitted to children's hospital neonatal intensive care units" addresses 625 cases of two hospital units, while P. Powell et al. analysed 762 newborn babies admitted to a single hospital [228] [229] [236] [240] [241] [242] [243].

Recent research works on LOS analysis can be found, such as "A Model-Based Approach to the Analysis of Patterns of Length of Stay in Institutional Long-Term Care" by H. Xie et al., "Grey correlation analysis of Average Length of Stay in General Hospitals" by H. Li et al., or the authors K, Mehmet et al. in "Cost Analysis on Intensive Care Unit Costs Based on the Length of Stay". All of these studies relied on statistical model analyses, which explored metrics such as basic statistical measures (minimum, maximum, mean, frequencies, …), statistic distributions, skewness, j-b statistics, correlation, among others [244] [245] [246].

Regarding LOS prediction is possible to find some research works, such as the one presented by Grampurohit et al. with the title "Hospital Length of Stay Prediction using Regression Models" which develops a comparation for LOS forecast based in four regression models (Linear, Ridge, Lasso and ElasticNet), the work developed by Kumar et al. described in the manuscript "A Two-Stage Model to Predict Surgical Patients' Lengths of Stay From an Electronic Patient Database" which clusters patients into similar LOS groups using the CART method or J. M. P. Gutiérrez et al. predicting LOS for several departments in a hospital, resorting to tree-like structures and radial basis function kernels described in their work "Predicting Length of Stay Across Hospital Departments" [227] [231] [247] [248] .

Despite the fact that these papers are recent, they do not use machine learning hyperparameters optimization automation (and in most cases, they do not even mention optimization), neither Neural Architecture Search nor eXplainable Artificial Intelligence (XAI) techniques, in their approach to LOS analysis - techniques that were developed in our research. Their methods are solely concerned with prediction, with no regard for the significance or influence of factors, which is a novelty introduced by this research.

A very recent and exhaustive systematic review on LOS prediction methods (published in 2022), reiterates the importance of LOS prediction as a valuable tool, confirms that most current research is conducted on adults who have some sort of pathology, and highlights the limitations of studies frequently conducted in restricted populations. Most importantly, it stresses the need for a standardised approach to LOS prediction and emphasizes the value of including XAI in the context of LOS research. The majority of the LOS shortcomings identified in this systematic review are, in our opinion, addressed by the research presented in this chapter [243].

When widening the literature review, we can understand that the use of XAI in health areas is still very limited and is mostly used to establish confidence in the generated models, rather than to reveal knowledge, as we suggest in the current study and the PhD thesis in general. When it comes to hyperparameter optimization automation and NAS work, they are also uncommon in the health field and the only examples are conducted on medical images [249] [250].

The lack of tools for NAS implementations, namely for NNs (which are crucial for this work, as we approach later), as well as some new regression optimization searches methods, have made it necessary for the thesis to consider new developments in this particular area. Part of these developments are documented in the CHAPTER 5 - Derivative Works (secondary outcomes) [166].

7.3 Business Understanding

The “business goal” for the use case scenario is to provide better information to all NICUs stakeholders so all parties can be aware of the factors which influence LOS at NICUs, as well as the dimension of the impact. The study outputs expect to provide a trigger for better investment options in LOS prevention at neonatal units or, to be more correct, to prevent or influence events/factors which increase LOS.

Considering the business goal, it was decided to develop a machine learning model that can accurately predict the LOS at NICUs, which means generate a model that can fit the data and modulate the relations between inputs and outputs. In another perspective a model able to embody the relationship between factors and LOS.

To obtain a performant model is the data mining goal. To be more precise, the objective also includes achieving the best possible accuracy, which will be attained by leveraging several techniques of architecture search and optimization techniques.

Once a performant model has been established, it is possible to analyse the features'/factors' importance and impact through the model's explicability (applying XAI techniques). This last analysis is the process business goal.

Because LOS prediction is a regression problem, the team decided to use the MAE - Mean Absolute Error complemented by the R^2 Score as evaluation/assessment metrics, following recent studies guidelines [251] [252].

The first metric, MAE, is expressed in the same unit as the target feature (LOS) which is expressed in days and provides information about the error range/dimension. The second considered metric is R^2 Score, which provides a general view of the model performance. This metric is not influenced by the unit's ranges and can be seen only as a relation. In this case, R^2 Score, will vary from 0%, meaning the model is completely unable to discover relations between inputs and outputs, and 100% representing a model perfectly able to express the relation between the inputs and outputs.

- MAE - Mean Absolute Error

$$\frac{1}{n}\sum_{i=1}^{n} \left|\hat{Y}_i - Y_i\right| \qquad \begin{array}{l} \hat{Y}_i \; predict \; value \\ Y_i \; label \; value \end{array} \tag{8.1}$$

- R^2 - R Squared

$$1 - \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - Y_i\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2} \qquad \begin{array}{l} \hat{Y}_i \; predict \; value \\ Y_i \; label \; value \\ \bar{Y} \; label \; value \; average \end{array} \tag{8.2}$$

7.4 Data Understanding

The source of data of this initiative is the output of a retrospective population-based study in which data was obtained from the first ten years of the National Registry of Infection in Portuguese NICUs. The original dataset is expressed by over 50k records which include all inborn and outborn neonates, of any BW and GA, admitted before the 28 days of life on all Portuguese NICUs between the years of 2008 and 2017. In our opinion, this is a very complete, comprehensive, and significative source of data which increases the importance and interest of the initiative outcomes. The previous studies in the area present quite limited data sources.

The initial data source encompasses the following information:

- 50161 admission records with the features: id, process id, hospital, birth date, admission date, discharge date, birth weight, gestational age, gender, surgical intervention, hospitalization days, hospitalization days in intensive care, TBI (the tracheal tube invasive ventilation), CVC (central venous catheter), UAC (umbilical artery catheter), NTISS (neonatal therapeutic intervention scoring system), CRIB (neonatal clinical risk index for babies), SNAP and comments.
- 10611 infection records with the features: id, admission id, infection type, start date (infection), death and origin (infection).

It is important to highlight that these presented features, the data collected from the retrospective population-based study, were previously discussed and approved by clinical experts in the scope of the "Programa de Prevenção e Controlo de Infeções e de Resistência aos Antimicrobianos" which can be translated as "Program for the Prevention and Control of Infections and Resistance to Antimicrobials" from the Portuguese Directorate General of Health [253].

Below it is described in detail each dataset feature, including the type of data and possible range:

Admission records

- Id (Number): unique identification of the newborn admission record;
- Process id (String): National Registry of Infection internal codification;
- Hospital (String): NICU name; The range includes all 19 Portuguese NICUs names;
- Birthdate (Date): newborn birthdate; Range 08-08-2007 to 29-12-2017;
- Admission date (Date): newborn admission date to NICU; Range 08-08-2007 to 30-12-2017;
- Discharge date (Date): NICU discharge date; Range 01-01-2008 to 31-12-2017;

- Birth weight (Continuous number): newborn weight at birth in grams; Range 300-5890 grams;
- Gestational age (Ordinal number): newborn gestational age in weeks; Range 22-42 weeks;
- Gender (Categorical): newborn gender; Range F, M or I (female, male or undefined);
- Surgical intervention (Boolean value): defines if the newborn was submitted to surgical intervention; Range True or False;
- Hospitalization days (Continuous number): newborn period of hospitalisation (or LOS) measured in days; Range 0-263 days;
- Hospitalization days IC (Continuous number): newborn period of hospitalisation in intensive care in days; Range 0-254 days;
- TBI (Continuous number): number of days which the newborn was submitted to the tracheal tube invasive ventilation; Range 0-242 days;
- CVC (Continuous number): number of days which the newborn was submitted to the central venous catheter; Range 0-208 days;
- UAC (Continuous number): number of days which the newborn was submitted to the umbilical artery catheter; Range 0-208 days;
- NTISS (Continuous number): neonatal therapeutic intervention scoring system; Range 1-82 or undefined;
- CRIB (Continuous number): neonatal clinical risk index for babies; Range 1-59 or undefined;
- SNAP (Continuous number): neonatal score for neonatal acute physiology; Range 1-99 or undefined;
- Comments (Text): comments and notes for the record.

Infection records

- Id (Number): unique identification of the record/infection record;
- Admission id (Number): relation to admission records; Range admission records Ids;
- Infection type (Categorical): infection type; Range Sepsis, Necrotising Enterocolitis – with or without sepsis, Sepsis with Meningitis, Sepsis with Pneumonia, Pneumonia, and Sepsis with Pneumonia and Meningitis
- Start date (Date): infection start date; Range 19-09-2007 to 25-12-2017;
- Death caused by infection (Boolean value): defines if the infection has resulted in the death; Range True or False;

- Origin (Categorical): defines the infectious origin. Range Maternal, Community, Hospital, Other Hospital.

As previously referred, the machine learning model developed for this study forecasts LOS (details in the following section 7.6 Modelling). In the dataset LOS is referred to as hospitalization days, which is the target variable.

The dataset was analysed from several perspectives such as features statistics metrics (e.g.: mean, variance), distribution, correlations between features and with the target variable, feature densities, among others.

More specifically, for each feature and for the target variable a complete data analysis was made, starting with counting records including empty, non-existing and non-valid values, data type, min, max, mean and standard deviations metrics, variance observations, sets of unique values, frequency counts and distribution, including outliers' detection and observation.

After the individual analysis, features relations, and features and target variable relations were analysed with the use of correlation metrics, trends between feature pairs, Cramer-V analysis for categorical features, chi-square tests or multivariate outliers. The goal of this analysis was to understand in detail the relation between features and/or target variable. E.g.: number of infections, by gender, year or month of admission, weight, or gestational age.

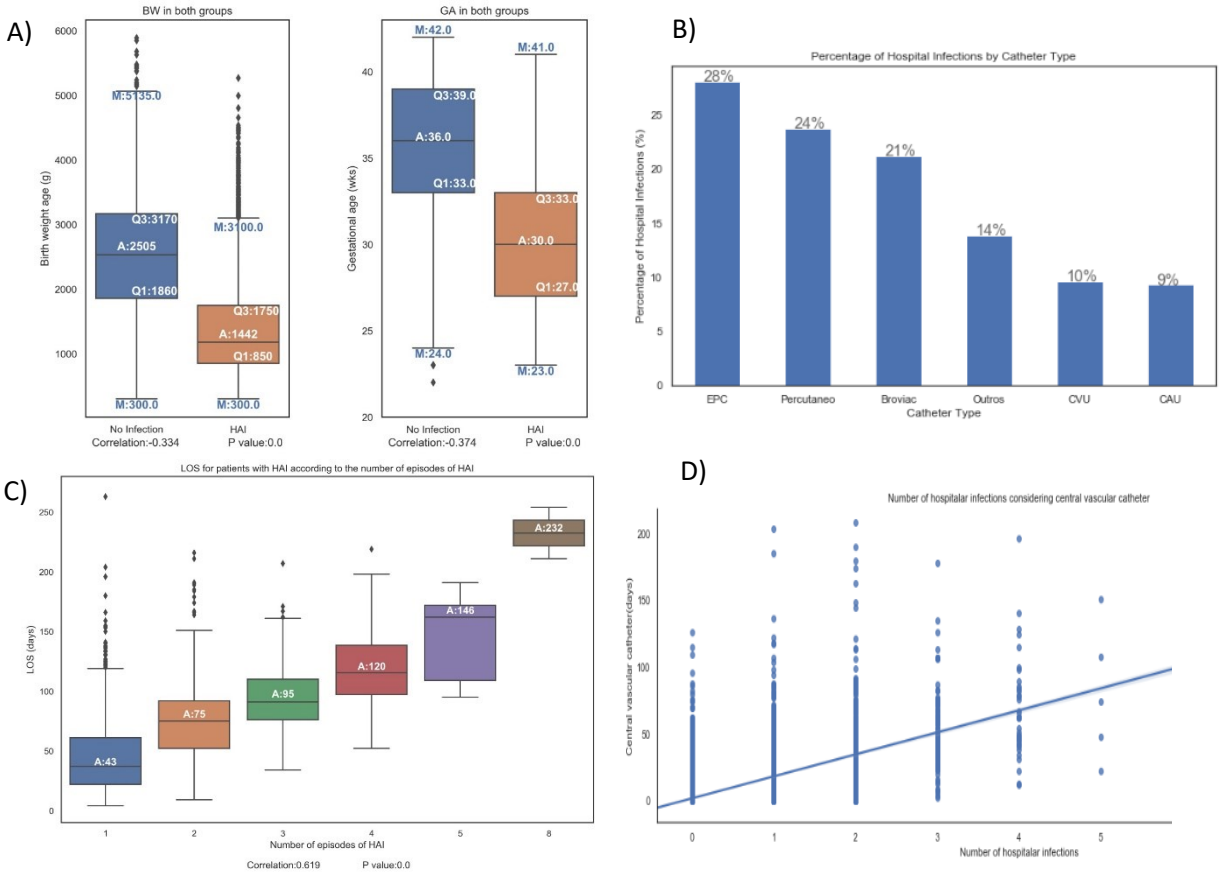


Figure 7.1 Example of plots created in the data exploration process

The analysis was supported by several visualization techniques from the simplest bar or histograms charts used, for instance, on categorical features counting and distributions observation, scatter plots for features relation analysis, line plots with errors bands for trends and error margins, heatmaps for correlation, boxplots, and violin plots for getting more insights on distribution and density, among others. Examples of the visualisation methods applied in this analysis are shown in Figure 7.1. It shows A) two boxplots that analyse the relationship between length of stay (LOS), birth weight, gestational age, and the presence of infection; B) a chart bar that offers some insights into how catheter use can contribute to infections; C) a view of the distribution of LOS that considers the number of infections for a single patient; D) a relationship between the number of catheter days and the number of infections.

It is important to emphasize that the target feature (LOS) presented several outliers (Figure 7.2, where the black diamonds represent the outliers), an additional challenge to data fit, which often negatively affects the model's performance. Nonetheless, because the outliers are considered clinically relevant, the clinical team involved in this study decided to include them.

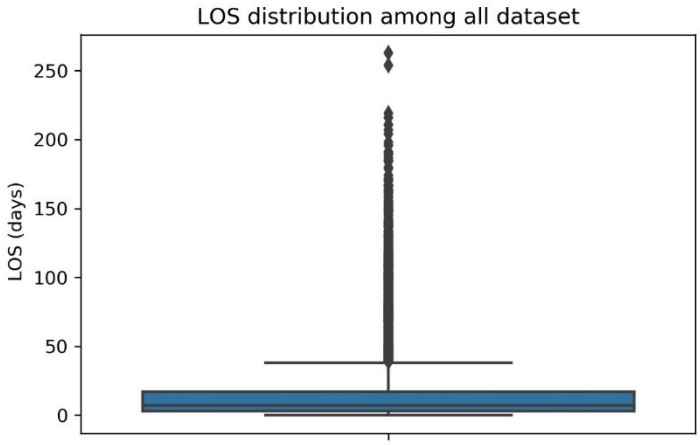


Figure 7.2 LOS distribution among all dataset (highlight for the incidence of outliers)

The LOS on intensive care could be a second variable to analyse but considering the current study purpose will be simply discarded. The feature was not considered for the model as, according to our analysis, constitutes a data leakage.

Also, the TBI, CVC and UAC days were signed by the analyses as data leakage information with a high correlation with the target variable and an obvious direct relation. E.g.: a patient with 40 TBI days has a LOS of at least 40 days. Considering the possible relevance of this information, it was decided to convert these features to Boolean features.

Verifications were made to discover incomplete or missing data, errors, duplicates, invalid data, and other data issues. For these verifications, several features were compared and related (e.g.: birth data > admission date). As expected on real data from real scenarios, several problematic situations were identified. For instance, considerable missing information on scores features, incompatible dates of birth and admission dates, episodes of infections starting before birth or after discharge, LOS on IC superior to LOS, among others.

Records including necrotizing enterocolitis (NEC), mother-related infections, community-acquired infections and infections acquired in other hospitals were excluded. The clinical team decided to exclude these infections records as well as associated admission records because of the clinical complexity of these conditions.

7.5 Data Preparation

The data preparation phase includes several tasks to reach the final dataset to model, which is crucial for the success of models' outcomes.

Following the previous data understanding phase, several features were excluded from admission records, namely unique id, hospital process id, LOS on IC, all scores, and textual observations. The following features were encoded: NICU name, gender (which presented three options) and surgical intervention from admission records, and death and infection origin on infection records.

Records with incomplete or missing data, or duplicates were excluded. Several errors were analysed in detail (e.g.: inversion of day and month on dates) and corrected when certain about the correct values, or when there was ambiguity discarded.

It was made some feature engineering creating derivative features as month and year from birth date and hospital admission date, days and weeks between birth and hospital admission, or the corrected age (a clinical concept) on hospital admission. The complete dates were excluded after this process as this type of data is not useful for the model. TBI, CVC and UAC were converted to Boolean features as defined in the previous phase.

NEC, mother-related, community-acquired, and other hospitals acquired infections and the associated admission records were removed, according to the instructions of the clinical team which integrated this initiative.

A final study dataset was created merging both admission and infections records and a new feature, namely infections number, was included in the final dataset. The final dataset is composed of 44 955 admission and 4 815 infections records.

The dataset prepared for the modelling phase was composed of the following features:

- BW;
- GA;
- Surgical intervention;
- Infections number;
- Death;
- NICU name (encoded);
- Birth year;
- Birth month;
- Hospital admission year;
- Hospital admission month;
- Birth to admission days;
- Birth to admission week;
- Gender M;
- Gender F;
- TET;

- CVC;
- CAU.

7.6 Modelling

The study challenge encompassed the need for developing a LOS forecast, which implies that the techniques used should output a continuous value, namely the number of days. Regression algorithms were considered for modelling the forecast.

Considering the transformations made on the data preparation phase, which have led to a dataset with exclusive numeric, Boolean, and encoded features, essentially, all existing regression techniques were available to be applied.

In the ML modelling process, the best practices demand to split the final dataset into training, validation, and testing data. One of the most common approaches was used, dividing the dataset on 70/20/10, respectively.

Considering the characteristics of the current regression problem, it was decided to first experiment modelling with the algorithms listed in Table 7-1, some of the most common used in ML research challenges. The results were produced using the default hyperparameters for each of the algorithms.

Table 7-1 Models performance for LOS prediction, using default hyperparameters

Algorithms	Training		Validation		Testing	
	MAE	R^2 Score	MAE	R^2 Score	MAE	R^2 Score
Linear	7.64	59.8%	7.70	61.2%	7.96	56.3%
Elastic Net	8.76	48.0%	8.82	49.2%	9.07	44.1%
Decision Tree	6.14	70.6%	6.64	64.2%	7.02	58.9%
AdaBoost	10.34	45.5%	10.37	46.3%	10.72	42.8%
Cross decomp	7.68	59.5%	7.77	60.6%	8.01	56.1%
XGB	6.35	68.0%	6.48	68.3%	6.87	61.6%

Table 7-1 presents the ML modelling preliminary results, using the algorithms default hyperparameters, which each open source library implements as the algorithms' "general version". XGB - eXtreme Gradient Boosting, a gradient boosted decision trees algorithm, performed well with a performance that attained for the training, validation, and testing datasets the following metrics: MAE (days) 6.35/6.48/6.87 and R^2 Score 68.03%/68.34%/61.55%.

After achieving a performance baseline, the goal is to improve the fitting model for the considered data, which is measured by MAE and R^2. Expert understanding of the algorithms to explore, including both algorithm concepts and algorithm training hyperparameters behaviours, is usually required for improving the model's performance (which is quite deep expertise knowledge). To obtain better results, the approach typically calls for manual regulation of hyperparameters, along with a lot of experiments, a dash of intuition, and some luck. In resume, it is an iterative, complex, and time-consuming process as already approached in chapter 4.1 Architecture Optimization .

For the task of trying to find the best training hyperparameters combination, it is common to use approaches such as random search or grid search for each algorithm.

Although random search or grid search are the common strategies, in this case it was decided to use the most recent innovations in the field, and automatic architecture optimization was used. The detailed process is described in the section that follows.

7.7 Architecture Optimization

Finding the best ML algorithm and the right combination of hypermeters for the chosen algorithm is a challenge within itself.

For instance, in a regression problem (like the one faced in this research) there are countless possible regression algorithms, such as linear regression, lasso, elastic-net, Bayesian regression, logistic regression, kernel ridge regression, support vector machines, nearest neighbours' regression, Gaussian Naive Bayes or decision trees regressions just to mention few. Each algorithm includes a unique and sometimes vast set of training hyperparameters with many possible values, which result in combinations are almost endless. This subject is better detailed in section 3.2.1 Architecture Optimization.

For the current phase it was decided to include in the optimization two main "branches" of machine learning, namely the "traditional" regression methods (which include, among others, the algorithms listed in the Table 7-1 and those described in the preceding paragraph), and the Neural Networks, namely multilayer perceptron NN.

Although "traditional" machine learning models are still frequently employed in today's world, it is also a fact that, considering technological advancements, NN approaches are increasingly being used to find the most efficient models. These approaches have a higher level of complexity, as they included two types of hypermeters, namely training and structural hyperparameters, but they also have a lot of potential for accurately fitting data.

Recall that among the training parameters are selections like learning rate, optimizers, and batch size. Choices regarding the number of units, type of units, activation functions, and many other options are included in structure parameters. Again, the possibilities are limitless.

The automatization for the best NN hyperparameters combination search was performed with the support of the concept Neural Architecture Search, a subject which is address with more detail in the chapter 5.1 Neural Architecture Search for (1D) CNNs. Note that the 5.1 chapter applies NAS to CNN while in this case multilayer perceptron NNs were applied.

One of the novelties offered by this work, as previously mentioned in the related work section, is the approach that is presented, which includes automation of hyperparameters optimization in the medical field, with highlight to NAS incorporating NNs on the search space.

The architecture optimization search for regression algorithms was performed for 72 hours, with a time limit for a model attempt of 30 minutes. The Bayesian search approach was used, with MAE minimization as the search objective (the optimization goal).

The NAS for NNs was performed using a maximum of 290 different attempts, a maximum of 1000 epochs, and an early stopping with patience equal to 30, to which has corresponded a process time of roughly 72 hours (4408 minutes). Time could not be used as a limit because NAS takes an entirely different strategy than the regression search. The NN search, like the prior regression algorithm search, was supported by the Bayesian search method and had MAE minimization as goal.

Both architecture optimizations were performed on the same machine and under the same conditions, namely:

- OS: Ubuntu 18.04;
- Disk: 500 GB SSD;
- Sys Mem: 32Gb;
- CPUs: AMD Ryzen 7 3600X;
- GPUs: NVIDIA GTX 1080 TI 11GB + RTX 2060 SUPER 8GB.

The obtained results are presented, below, in Table 7-2.

Table 7-2 Architecture optimization best results for regression algorithms and NNs

Algorithms	**Training**		**Validation**		**Testing**	
	MAE	**R^2 Score**	**MAE**	**R^2 Score**	**MAE**	**R^2 Score**
Regression Alg	5.41	78.2%	6.14	69.5%	6.55	63.5%
NN	5.88	66.9%	6.28	69.4%	6.83	60.2%

The best model for regression algorithms, and the best model overall, was achieved with the Light Gradient Boosted Machine (LGBM or LightGBM) Regressor algorithm, using as parameters (only the most relevant are presented): boosting_type=gbdt; learning_rate=0.02108058601944604; max_depth=174; n_estimators=687; num_leaves=44; reg_alpha=2.0957249919349216e-05; reg_lambda=0.00045050730969185853.

LGBM Regressor is an open-source method designed by Microsoft that implements an efficient and effective gradient boosting algorithm that is supported by tree-based learning algorithms. The feature that distinguishes it from other boosting algorithms is the automatic feature selection and a focus on boosting examples with larger gradients. These distinguishing characteristics "promise" faster training and better performance.

In practice, the algorithm uses discrete beans to sort all features and bucket the observations, focusing on data samples with larger gradients. When deciding on a new tree leaf to create, the algorithm uses previously created buckets in a leaf wise strategy (meaning grown depth first), rather than the common level wise strategy used by other boosting methods. This strategy is also referred to as histogram based [254] [255].

Although the huge hype around NN, the current challenge has proven to be best suited with a gradient boosting algorithm, in this case Light GBM. This is no exception as boosting algorithms have presented interesting results on other types of challenges.

Despite the clinical team's decision to keep the outlier's records on modelling process because they are considered clinically relevant, it was decided to build another model using LGBM Regressor trained with outlier exclusion just for complementary information (2068 training records excluded). The outcomes are shown in Table 7-3:

Table 7-3 LGBM regressor algorithm performance excluding outliers' records

Algorithms	Training		Validation		Testing	
	MAE	R^2 Score	MAE	R^2 Score	MAE	R^2 Score
LGBMReg	4.23	84.8%	4.82	74.8%	4.93	76.6%

7.8 Evaluation

The previous architecture optimization phase was successful in producing a good model performance with a 78% accuracy (measured by R^2 Score) on the training data set, almost 70% on validation and 63% on testing. The MAE was between 5,41 and 6,5 days (Table 7-2).

The results of the model training with the exclusion of outliers showed a positive impact on model performance, which was thought to be likely (still) during the data understanding phase. The three sub-datasets performed better overall thanks to the exclusion of outliers, with an improvement in accuracy of up to 13% and a reduction in MAE of about a day. The testing dataset, which is the set responsible for capturing the model's capacity to generalise the forecast, was the sub-dataset that showed the highest gains.

Given the current real data and the challenging real scenario, the performance is regarded as being remarkably interesting from a technical standpoint. From a business perspective, and considering the models discovered, the current performance is distinguished by a trustworthy and reliable LOS forecasting, reaching a (maximum) R2 Score of 85%.

Due to the units used and the experiment setup, which is unique in LOS, the results obtained in this research are not comparable to those obtained in SotA studies which rely on statistical analysis.

Despite the significance of the model performance, the primary objective of the study was to identify and quantify the factors that affect LOS at NICUs. The following chapter provides more information on the explanation techniques that were employed.

7.9 Explanation

Model explanation is performed by resorting to the so-called XAI techniques. In this instance, Shapley Additive Explanations techniques were selected as the best option to explore which features most significantly affect LOS and the extension of those effects.

In a resumed way, Shapley Additive explanations, also known simply as SHAP, were introduced in 2016 by Lundberg and Lee. These techniques are based on a game theoretic approach and are used to explain ML predictions, namely by computing the contribution of each feature to the prediction [136].

These techniques are approached in detail in the section 4.2.7 SHapley Additive exPlanations (SHAP).

In the Figure 7.3, produced with the support of SHAP library, it is possible to visualize a LOS prediction for a particular admission record with HAI, along with the corresponding explanation. Or to put it another way, it is an explanation of a specific LOS prediction instance.

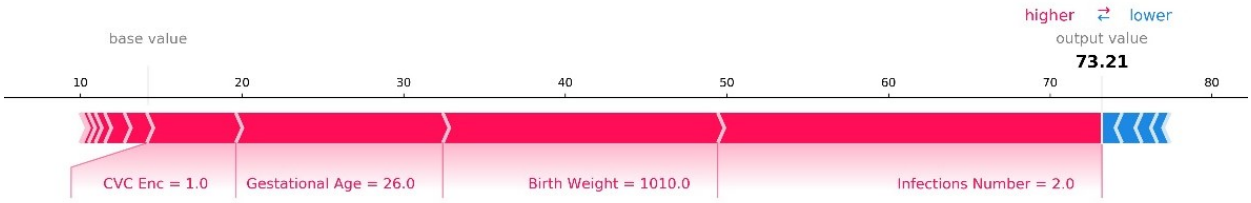


Figure 7.3 Features impact on LOS prediction (record with HAI)

The figure makes it possible gather a plethora of information. It is possible to observe that the model produces as output a LOS prediction of 73,21 days for the specific record/data instance. The colour red highlight the characteristics that increased the LOS value and show the corresponding influence. It is possible to see that this patient experienced two episodes of HAI (feature infections number with value 2), which increased LOS in about 22 days. The patient's birth weight of only 1010gr is the second characteristic that mainly raises LOS. The increase is around 17 days. GA valued as 26 weeks accounts for 13 additional days, and so on for the other features. The impact of these contributions is always ranked, and the contributions with the greatest impact are displayed first (next to the prediction output).

Contributions that reduce the output prediction are represented by the blue colour. For this example, the effect of these features can be considered residual.

This Figure 7.3 provides a good idea of this use case goal: to identify the variables that affected LOS and their corresponding effects, in this case observing a single instance. All records in the dataset can be subjected to the same previous analysis. This approach is a novelty in the LOS analysis field.

Using Figure 7.4, another tool provided by SHAP, it is possible to examine the contribution of all examined factors to the LOS prediction from a different angle. The graph does not provide a forecast for the specific instance, but it does offer a thorough overview of how all features contribution for the prediction, defining the type of influence (positive or negative) and the corresponding impact.

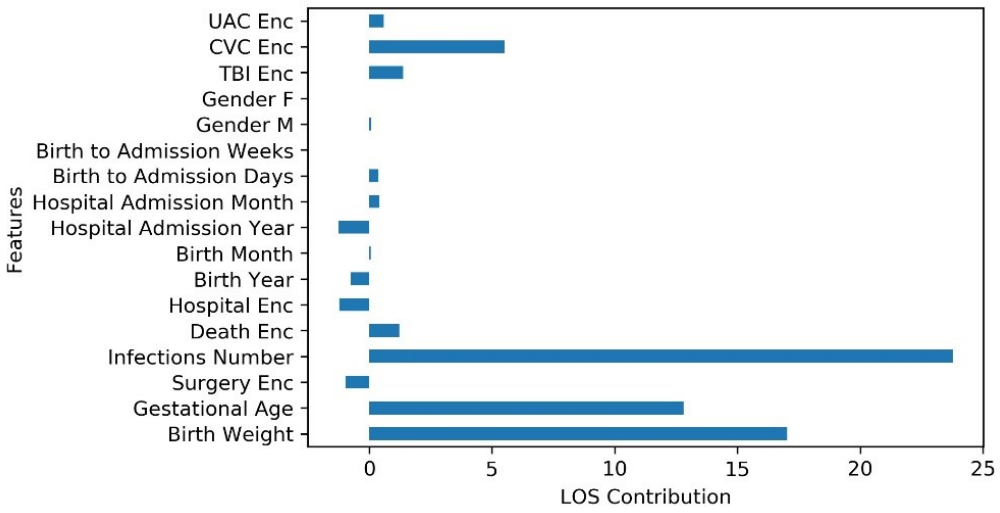


Figure 7.4 Features impact on LOS prediction (record with HAI); another perspective

Another example of LOS prediction is shown in Figure 7.5, but in this case for a record without HAI (which roughly contrasts with Figure 7.3).

The factor that has the greatest positive impact on LOS prediction (lowering the model outcome) is the birth weight of 3668 gr. The second most significant factor is the gestational age of 40 weeks. Both factors are considered healthy at birth, and their contribution to LOS expectancy was expected. Finally, the third factor is not having any HAI (in the plot is possible to verify that infections number = 0). The features that follow can be read in a similar manner.



Figure 7.5 Features impact on LOS prediction (record without HAI)

After exploration and understanding of several prediction examples, it is possible to move on to an all-population analysis to extract more comprehensive insights. Plot A, presented in the Figure 7.6, explains the influence of each factor on LOS, ordered in decreasing magnitude. BW is the factor that most influences LOS, followed by GA and HAI which occupies the third position. The second plot (plot B) provides another dimension perspective, namely the impact of the factors. The plot indicates that the infections number feature has the greater impact on LOS, with an effect that can reach more than 70 days. This factor clearly stands out from every other factor. Later in the document, a more thorough analysis of the feature allows for a more in-depth understanding of this behaviour.

Following the two plots presented in Figure 7.6, it is clear to state that the three most influencing and most impactful factors are HAI, birth weight and gestational age.

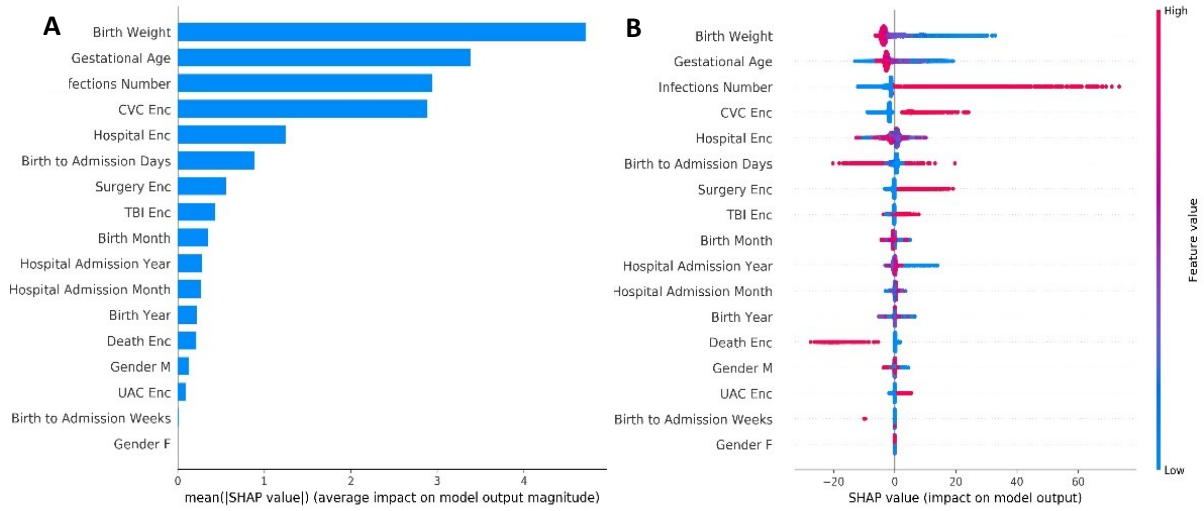


Figure 7.6 Factors influence and impact on LOS

A call for attention to the death feature (Boolean encoded) which must be read carefully. With an inattentive reading, the analyst could be tempted to claim that higher values - meaning death has occurred - have a positive impact with significant LOS reduction; and lower values - meaning non-death - would have a neutral effect or even a small contribution to LOS increase. Death is undoubtedly the worst scenario and the clinical first goal to avoid. The only reason for what the plot shows is the termination of the admission time due to premature death. Premature death is even worse than extending LOS.

After analysing several features from a global perspective, the knowledge can be expanded by individual feature' analysis. The analysis of the three most important features is shown in Figure 7.7. Be aware that each plot's horizontal axis contains different variables and scales.

Focusing on the vertical axis for each plot - which presents the LOS contribution in days - it is possible to understand that the axis limits are quite different, further emphasising how different each feature's impact dimension is.

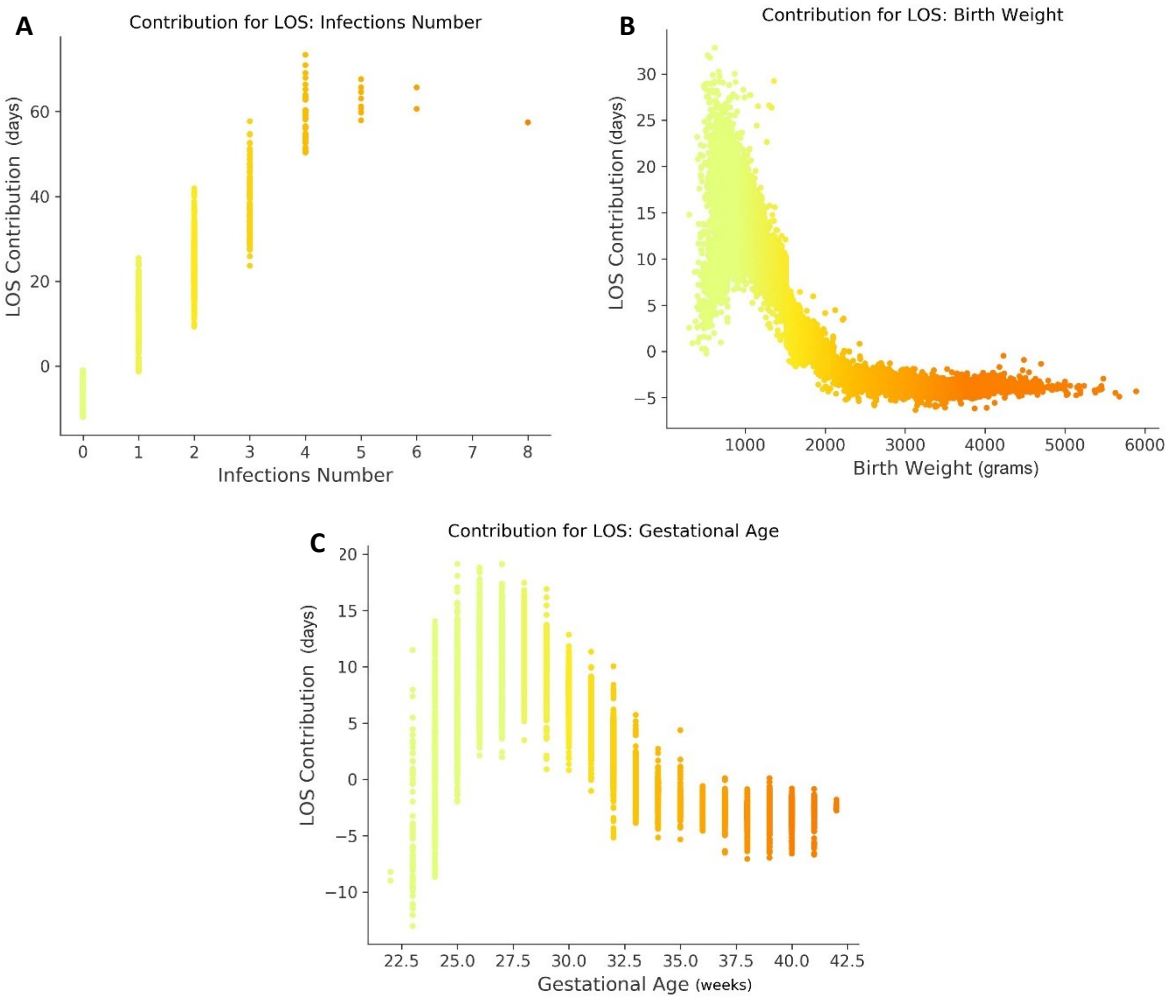


Figure 7.7 Detail analysis for the three most relevant features for LOS: a) infections number cases; b) birth weight; c) gestational age

The absence of HAI has a beneficial effect on LOS, reducing it. With an increase in infections, the LOS significantly rises (Figure 7.7. plot A).

Regarding birth weight, it is possible to observe that a birth weight greater than 2100 gr usually contributes to the LOS reduction (observed in the LOS negative values). From 3000 g on, the incidence becomes more pronounced (up to less 6 days). Values below 2000 g cause the LOS to increase by up to 31 days (Figure 7.7. plot B).

With a LOS reduction contribution for GA>35 weeks (LOS contribution negative values), GA behaves similarly to BW. Figure 7.7. plot C, shows that below this point, the LOS contribution gradually increases up to 19 days.

Be aware that both last features, on the lowest values (for instance 1000gr for birth weight) may be affected by records that resulted in death and that they may appear to positively contribute to LOS reduction, which can lead to misunderstandings.

7.10 Deployment

The current initiative incorporates the objective of developing a machine learning model for LOS prediction, but not with the purpose of making the LOS forecast accessible to professionals, even though that could be also possible with the achieved results.

The LOS prediction model goal was developed as a tool for learning more about the factors influencing LOS, as well as the dimension of the impact. XAI techniques helped to support the acquisition of these insights. As a result, this initiative won't lead to the deployment of an IT solution, but rather to research findings, which will be presented in the next chapter.

These results have as final recipients all NICUs stakeholders and aims to contribute to improve investment options in LOS reduction at neonatal units.

7.11 Research Results

The purpose of this study was to comprehend how the factors that were analysed influence and affect LOS, using NICUs as a use case.

It was possible to rank all the factors according to both influence and impact, as was seen in the explanation chapter. The most distinguishing factor was HAI, in which it was also possible to comprehend the impact of the number of episodes of infections. To the best of our knowledge, no studies have been conducted on the effect of the number of infections in LOS, and those it is a novelty introduced by this study.

Figure 7.8 presents on the right side the LOS average categorized by the number of episodes of HAI, while the boxplots show the specific contribution of the number of episodes to the total LOS. Just a single infection episode can increase LOS prediction, on average, in 11 days, and can go up to 22 days. Having two episodes increases the average contribution to LOS in 28 days, which is more than twice as much as with just one. Looking at the overall picture, the effect of the number of infection episodes is easily understood; one additional episode corresponds to roughly 15 additional LOS days. It is worth emphasizing that records with 5 or 8 infections are uncommon in the study dataset. The small number of records makes it impossible to extract consistent information.

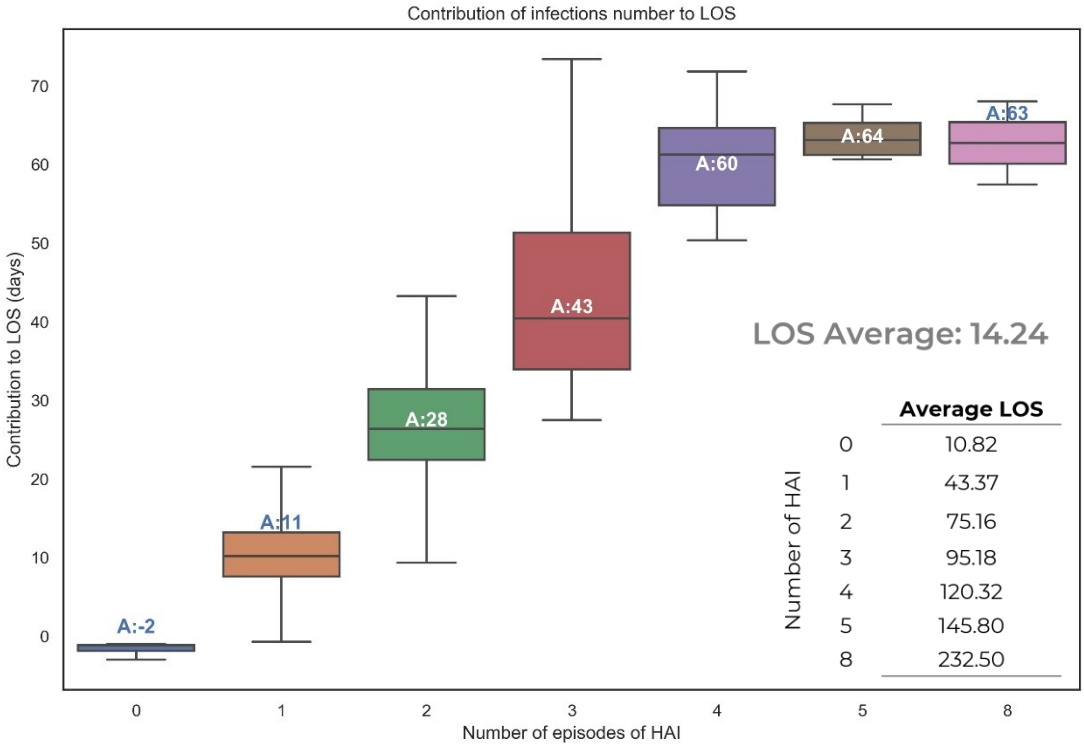


Figure 7.8 Impact on LOS by number of episodes of HAI and LOS total duration

It is also possible to observe that the average length of stay (LOS) for records with a single HAI episode is 43 days, which means that the infection is responsible for 25% of the LOS (11 days). In a two-episode situation, the mean LOS is 75 days, and infections now account for 37% (28 days) of the overall LOS.

Considering the population with HAI (one or more episodes) to which corresponds 3870 records, we found an increase of 59.173 hospitalisation days for the entire population. In contrast, the absence of any HAI (when the total number of HAI episodes is zero) led to a positive LOS reduction, which, according to Figure 7.8, results in an average LOS reduction of 2 days. When all 41085 records without HAI and the corresponding LOS reduction are taken into account, the cumulative reduction effect is 73.435 days.

BW was the second factor most relevant to LOS prediction. According to most of the neonatal literature the BW can be classified on extreme low birth weight (BW<1000g), very-low birth weight infants (BW<1500g), low birth weight infants (LBW<2500g) and newborns with BW greater than or equal to 2500g.

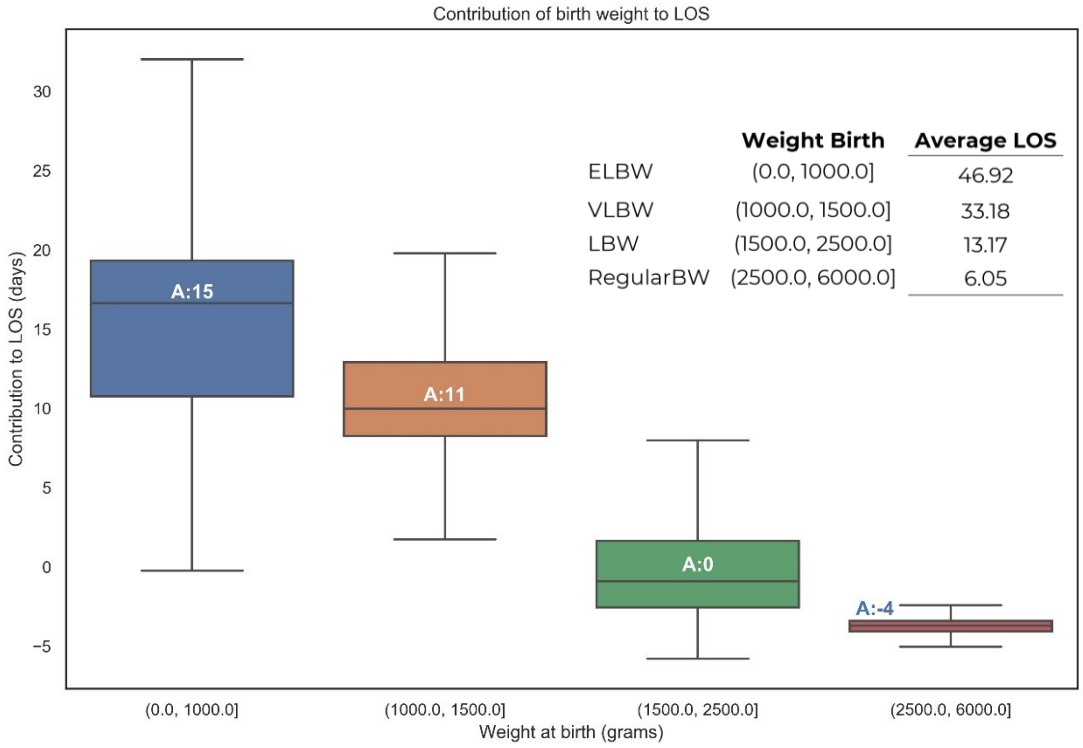


Figure 7.9 Impact of each BW category on LOS

The contribution of each BW category to the overall LOS prediction is shown on Figure 7.9, right side. It is evident that lower BW causes an increase in LOS prediction, with an average increase of 4 days when comparing a BW of 2500 and LBW, and an increase of 11 days when considering LBW and VLBW, which is nearly three times the increase between the previous categories with only a 500gr difference. ELBW continues to show an increase in LOS expectancy. The increased range is affected by premature death records with less LOS, and by few records to analyse. Given these restrictions, the value must be carefully analysed.

From the above information, it is also possible to extract an influence of 33% on total LOS for VLBW and at least 34% for ELBW.

Regarding GA contribution to LOS, it is possible to have a wide perspective on the impact of the factor weeks of gestation at birth through Figure 7.10. In general, births on the 34th week, do not affect the LOS prediction, and starting with the 35th week onward the contribution for all records is positive and help to reduce the LOS prediction by an average of 2.5 days.

Births occurring before week 33 gradually extend the expected LOS by up to 19 days. On average, week 33 contributes 0.61 days, week 32 contributes 2.95 days, week 31 contributes 5.52 days, and so on. On average, a one-week reduction in GA results in an increase in the LOS prediction in 2.5 days, but the valued tends to rise when the GA decreases.

To the best of our knowledge, this is the first LOS study to offer a detail view for every week of gestation.

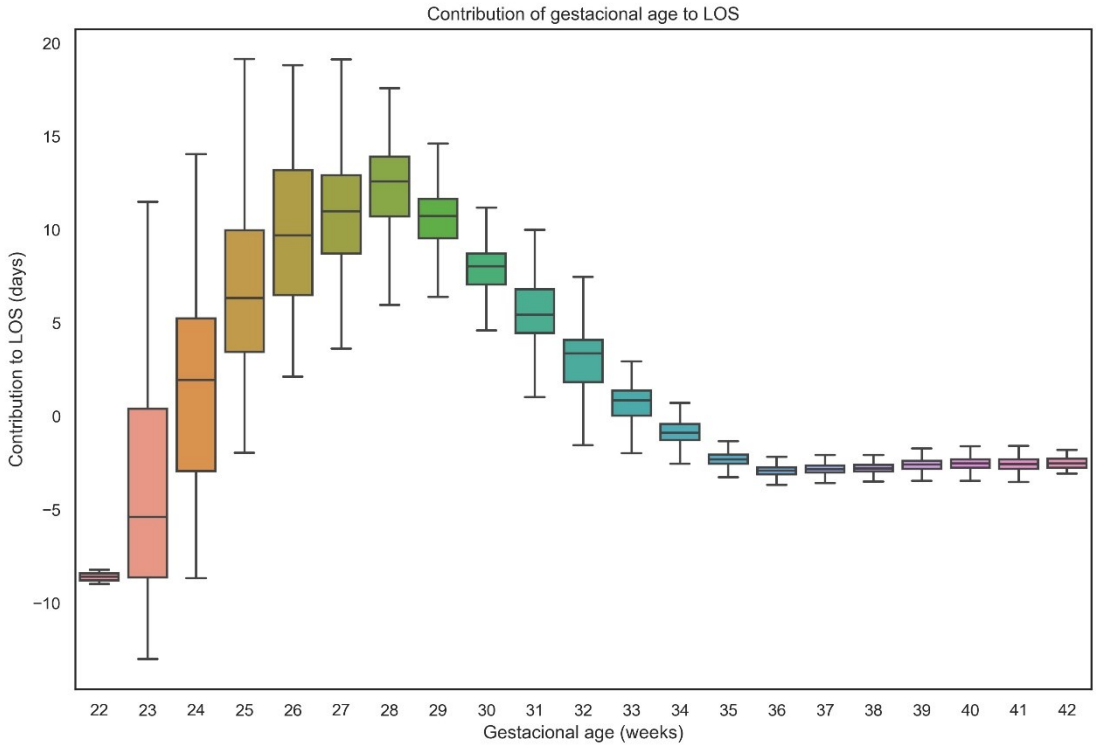


Figure 7.10 Impact of GA on LOS

The lowest GA values, as was already mentioned, have fewer records and are where we find more premature deaths. These two factors lead to the left skewed trend shown in the Figure 7.10. When drawing insights and conclusions for this GA range, that must be considered.

7.12 Conclusions

The proposed study of LOS at NICUs, resorting to a new approach based on data science techniques, instead of a "traditional" (and SotA) statistical analysis. To the best of our knowledge, LOS was analysed in a multidimensional way considering relevance and impact, for the first time. This proposed method approaching length of stay analysis at neonatal care units with data science was published in an IEEE conference [2].

With the help of this new approach, medical professionals and decision-makers may be better able to make choices that may reduce the incidence of events that affect LOS and the effects of LOS on newborns. The reduction of cost with additional medical care and the saving on scarce resources, are other benefits.

The study was able to produce a performance LOS prediction machine learning model, which allowed, through XAI techniques, to analyse factors/features contribution for LOS. The outcomes of the research process allowed for the confirmation of SotA claims regarding the significance of gestational age and birth weight for LOS as well as the clarification of the relevance of HAI. Furthermore, the study has added an innovation that gives LOS analysis a new perspective: the perception of feature relevance and impact.

The analysis clearly identifies three main factors that are relevant for LOS at NICUs namely and order by relevance, HAI, BW and GA, as well it delivers concrete measures of impact, such as the impact of having additional infection episodes or less one gestational week. We are aware of no other LOS study that offers this level of detail.

For instance, having no infections contributes to an average decrease in the LOS model prediction, whereas (just) one episode of infection results in an average increase in LOS of 11 days, and a maximum increase which can reach up to +20days; two episodes of infection can result as much as +40 days, and so on.

The problem isn't the additional in-hospital days per se, but also the new courses of invasive ventilation, new central lines, antibiotics, additional exams, to enumerate only a few extra needs. These needs contribute to add complexity to the clinical scenario, which has an impact on patient health and depletes scarce resources.

The decrease of BW and GA stems from technological advances in health neonatal care and better chances of survival in not so perfect conditions, and those factors must be managed. In opposition, HAI is a factor that may get more attention. It is possible to define and implement programs and new processes that directly address the identified problem with a likelihood of return.

Recall that according to our analysis it was estimated an aggregate impact of HAI on LOS of 132.608 days when considering the population under study.

The research team covets that the current initiative “direct” outputs may contribute to improve decisions on neonatal care and advance the general health of our society.

The research also resulted in a new model for NICU LOS prediction. Its use in a practical environment may be considered in a near future.

The research was conducted using a dataset which incorporates a quite interesting and meaningful population, but additional studies using the suggested technique and/or in other populations are welcome and can contribute to the enrichment of the knowledge. For example, it would make possible to analyse outcome differences and attempt to understand their origin by applying the same technique to a population in a different country. These differences can be related to several factors, such as various clinical procedures or patients care policies.

The proposed data science technique was applied in this case to analyse the length of stay (LOS) at the NICU, but it can also be used to analyse other LOS, clinical domains, and health care challenges, like discovering the factors that affect a particular surgery's success.

The research team wishes to encourage other multidisciplinary teams to approach their health challenges from a fresh data science perspective.

CHAPTER 8 - Applied Data Science: Analyse of Cardiovascular Changes Due to Obesity in Children

The increasing prevalence of overweight and obesity is a worldwide problem, with several well-known consequences that might start to develop early in life, during childhood.

The present research, based on data from children that have been followed since birth in a previously established cohort study (Generation XXI, Porto-Portugal), takes advantage of SotA data science techniques and methods, including Neural Architecture Search, eXplainable Artificial Intelligence, and Deep Learning, to explore hidden value in data, namely on electrocardiogram records performed during follow-up visits.

The combination of these techniques allowed to clarify subtle cardiovascular changes, already present at 10 years of age, evident from ECG analysis, induced by the presence of obesity. The proposed novelty combination of new methodologies and techniques is discussed, as well as their applicability in other health domains.

8.1 Introduction

Paediatric obesity, in both children and adolescents, constitutes a major global public-health problem. Over the last decades and according to the 2020 World Health Organization (WHO) report, the prevalence of paediatric overweight and obesity rose dramatically from 4% in 1975 to 18% in 2016 (over 360 million children and adolescents). The worldwide obesity has nearly tripled in 2020 when compared to 1975 [256] [257] [258].

World Health Organization (WHO) defines overweight and obesity as abnormal or excessive fat accumulation that presents a risk to health; in adults, a body mass index (BMI) of 25kg/m² or higher is considered overweight, and a BMI of 30kg/m² or higher is considered obese. BMI is a straightforward measure of weight-for-height that is frequently used to categorize individuals as overweight or obese. It is calculated by dividing a person's weight in kilograms by the square of his height in meters (kg/m2). Example: Weight = 68 kg, Height = 165 cm (1.65 m) results on the following equation: BMI calculation = $68 \div (1.65)^2 = 24.98$ [259] [260].

From birth to adulthood, individuals are susceptible to multiple influences, which can determine a higher likelihood to become overweight or obese. There is evidence supporting that metabolic and cardiovascular (CV) consequences of obesity, such as hypertension, dyslipidaemia and endothelial function impairment, and the development and progression of atherosclerosis, begin early in life and accompany the obese child throughout the life [261] [262] [263] [264] [265].

The clustering of CV risk factors in early childhood is concerning, given that ~80% of obese youth remain obese in adulthood. Recently, several environmental changes have enlarged the rise in the prevalence of childhood obesity and created the vicious cycle of childhood obesity. The root causes of obesity development in childhood and adolescence reflect complex interactions among environmental, socioeconomic, behavioural, and genetic factors [256] [266].

There is solid evidence supporting obesity as an independent risk factor for cardiovascular disease (CVD) in adults, but the evidence is scarce in paediatrics. Nonetheless, obesity and its risk factors do not influence all individuals equally. Thus, prevention, screening and follow-up become a challenge for all physicians and researchers [267].

There are several studies about obesity, but few of them include children. These studies typically focus on risks, causes, prevention strategies, connections to diseases, behaviours, lifestyle choices, and other factors. The UK-based Avon longitudinal study of parents and children, which analysis the strongest factors associated with the obesity risk in a cohort of >8,000 children, is a good example of a study that includes children [268].

The current study has a different focus than earlier research as: first, it examines effective cardiovascular changes brought on by obesity; and second, it is conducted on children, a population that typically receives less attention on this subject.

It is known that obesity increases the risk of cardiovascular disease, but does it cause cardiovascular changes in children age? If so, what exactly are they?

The current initiative sets as its objective to uncover the answers to these questions while deepening understanding of obesity and its consequences, namely on children.

8.2 Related Work

As mentioned earlier, obesity is a worldwide problem, and childhood obesity contributes to the issue (at the very least as a factor in future adult obesity). However, there are also reports of harmful health effects throughout the life course linked to childhood and adolescent obesity [269].

The cardiovascular function assessment, in children and adolescents, may include a standard 12-lead electrocardiogram (ECG), pulse wave velocity analysis, blood pressure measurement and 24-hour ambulatory blood pressure monitoring [270] [271].

Studies performed on adult population present evidence suggesting a close association between obesity and cardiac abnormalities. The abnormalizes are highlighted at the ECG wave level, including leftward shifts in electrocardiographic axes, left ventricular hypertrophy, and flattening of the T wave, which can be identified by an experienced cardiologist. Moreover, some of these ECG changes may be reversed with weight loss. However, the correlation between obesity and cardiac abnormalities in children and adolescents has not been studied in detail and hardly any concerning the correlations between body fat distribution and ECG [272] [273].

In general, epidemiological research on obesity impact are typically (if not exclusively) produced by standard statistical analysis, where the output generally includes multiple linear regression analysis, ratios, correlations, among other statistical measures. On one hand, these studies mainly address the risk factors that lead to obesity, namely socioeconomical status, lifestyle habits, sleeping patterns and genetic influence. On the other hand, they may also evaluate the possibility of an association between anthropometric parameters, such as body mass index, waist-to-height ratio or waist circumference, and the prevalence of cardiovascular disease (CVD), insulin resistance, chronic kidney disease or dyslipidaemia [267] [274] [268].

Some of these studies confirm the role of increased BMI on early markers of CVD related to childhood obesity. One example is the study performed by Twig et al. which analyses an Israeli cohort relating BMI at age of 6-19 years and claims a correlation between death and CVD later in adulthood [275] [276].

Overall, the research is focused on adulthood and long-term analysis of paediatric obesity impact. The scientific data on CV changes throughout childhood and adolescence is scarce. This initiative main aim is to address cardiovascular changes in children due to obesity, using electrocardiogram (ECG) analysis exploration supported by data science techniques.

8.3 Business Understanding

The current initiative aims to assess the existence of cardiovascular changes in obese children, aged 9 to 11 years old, identified by ECG analysis exploration with data science techniques.

With this research, it is expected to learn more about early cardiovascular markers that can be identified from an early age, namely from childhood, and therefore to minimize the obesity impact later in adulthood.

To achieve the proposed objectives, the research team worked in collaboration with the Generation XXI project, a previously established prospective population-based birth cohort from Porto - Portugal that aimed to characterized prenatal and postnatal development throughout childhood, adolescence, and adulthood. This birth-cohort was established in a well-delimited geographic area in north of Portugal, with a total of 8647 children, consecutively born alive between April 2005 and August 2006 in one of all five public maternity units covering six municipalities of the metropolitan area of Porto. Throughout the sequential evaluation, all participants were submitted to anthropometric, laboratory and cardiovascular assessment (including blood pressure and ECGs). The project gathered the ECGs recorded during the ten years' evaluation, and this research uses data science techniques to explore ECG changes [277].

Table 8-1 Generation XXI cohort

Characteristics	**Description**
Cohort recruitment date	Abril 2005 – August 2006
Localization	Porto metropolitan area
Hospital units	Centro Hospitalar de Vila Nova de Gaia/Espinho, Centro Hospitalar do Porto - Maternidade de Júlio Dinis, Hospital de São João, Centro Hospitalar do Porto - Hospital de Santo António e Unidade Local de Saúde de Matosinhos – Hospital Pedro Hispano
Population	8647 children, born alive between April 2005 and August 2006 with at least 24 weeks of gestation
Data collection process	Face-to-face interviews and diagnostic tests conducted in the hospital units and at the Generation XXI project facilities
Data collection periods	Birth, 6, 15 and 24 months, and at 4,7, 10 e 13 years old (still ongoing)
Data collected	Various health-related parameters (social, behavioural, organisational, biological), in order to monitor population evolution. E.g.: anthropometry and blood pressure measurements, blood sampling and bioimpedance, electrocardiogram, spirometry, pubertal status evaluation and cognitive evaluation

The strategy for the data mining process consists of developing a performant machine learning model, able to classify the ECG beats in obese and in normal weight children. With the model creation phase completed, it would be possible to analyse the model "hidden" structure and explore the model classification basis. This analysis will be performed with eXplainable Artificial Intelligence techniques.

In the research context an ECG beat or simply heartbeats refer to a complete PQRST complex interval, which is basically a form of characterization of a human heartbeat. In a more technical way, the QRS wave is produced by the atrioventricular node and is responsible for ventricular depolarization. T wave maps the ventricular repolarization and P wave indicates atrial depolarization [278].

Taking into consideration that classifying heartbeats, as from normal weight or obese children, is a machine learning classification problem, and also the data mining goal (for this use case), several metrics can be used for measuring the model performance. Some of these include a confusion matrix and the associated measures (e.g.: accuracy), receiver operating characteristic curve (ROC), cross-entropy, among others. The choice of the measure depends highly on the problem, data, and on the objectives to achieve. Considering all research variables and aims, it was decided to use accuracy as the evaluation/assessment metric.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negative}}{\text{True Positives} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \quad (8.1)$$

Although the data mining objective is to create a performant ML model that can distinguish between obese and normal ECG beats, this is "just" required step to extract the knowledge and insights needed to achieve the research objective, which is the identification of potential cardiovascular changes brought on by obesity.

8.4 Data Understanding

The data source of this study was directly extracted from an ECG equipment that was assigned to the Generation XXI project. The equipment was not used exclusively for this project.

The extraction of ECGs has processed 8224 valid records in XML format, on a proprietary vendor format.

The information contained in the XML file was extracted to a more "friendly" format, making it more suitable to the data mining process. This procedure was accomplished using the python programming language supported by the "Beautiful Soup" open-source library [279].

The following information was extracted from the XML format:

- ECG collection date
- Individual (related data)

- Identification (encoded)
- Birthdate
- Weight
- Height
- Gender

- ECG signal (Lead I)
- ECG report

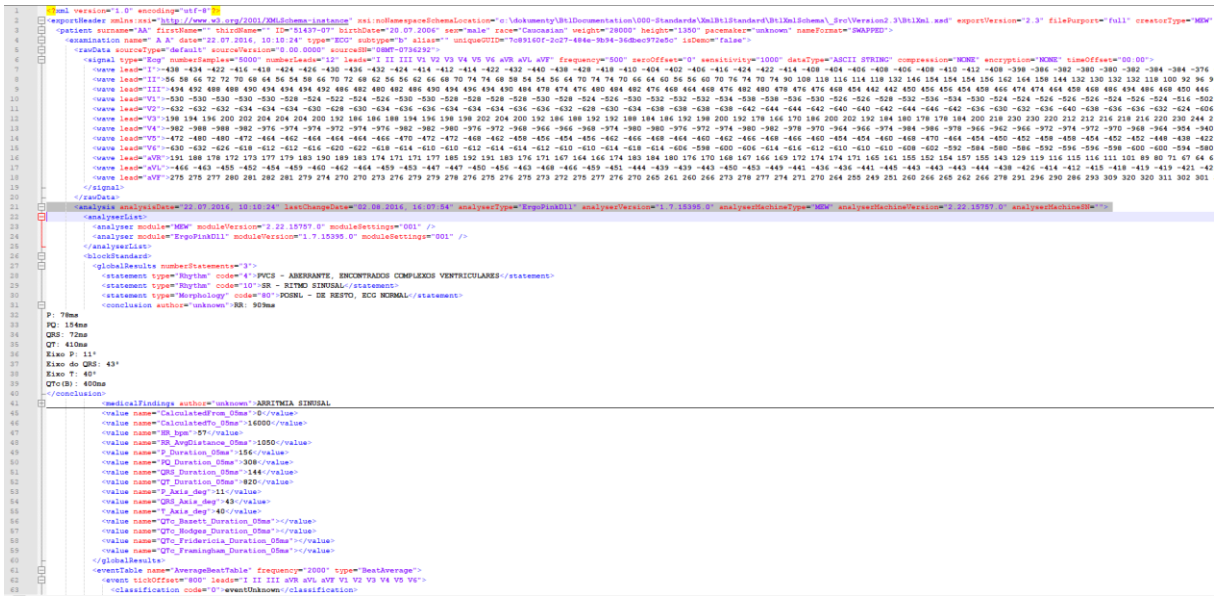


Figure 8.1 An example of an extracted XML file after anonymization

Since the ECG equipment was not only used in the Generation XXI project, data extraction from individuals who weren't part of the Generation XXI cohort was a possibility. The first step was to analyse ECG records considering the individual age at the time the ECG was collected. The age distribution records are depicted in the Figure 8.2.

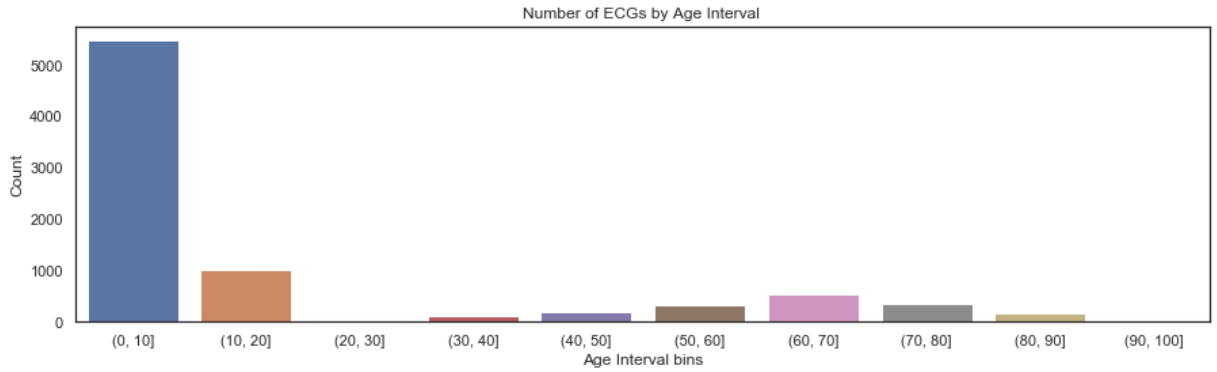


Figure 8.2 Distribution of initial ECG record extraction, by age classes.

The Generation XXI ECG collection was performed during the cohort follow-up visit at 10 years of age, but some children were evaluated a few months before or after the age of 10 (with a few already having completed 11 years of age). Thus, all ECG collections between the age of 9 and 11 were considered for further analysis.

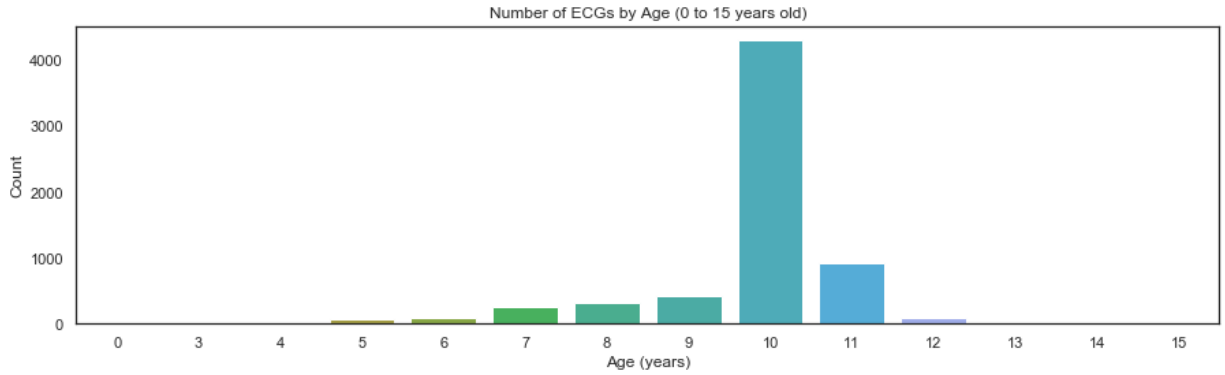


Figure 8.3 Distribution of initial ECG record extraction, between 0 and 15 years of age.

The extracted records binned by age are shown in Figure 8.3, with the age range of 0 to 15 years of age as filter. We can observe that some ECG records belong to individuals in ages different from our target population. These are records from children whose ages are not between 9 and 11 and who were certainly not a part of the cohort.

In a way to assure the proper treatment during the following data mining phase, a complete anonymized ID list of all Generation XXI participants was used and merged with the existing ECG extracted records, to guarantee the inclusion of only Generation XXI cohort participants. The final collected dataset included 4306 records from children with 10 years of age, 917 from children with 11 years of age, and 423 from children with 9 years of age. Following the aforementioned actions, the data collection task was finished.

Signal extraction from the ECG records presented a sampling rate of 500 samples/s, with 10 seconds duration, which results in a total of 5000 samples for each record.

All ECG records were “scanned” to find duplicates, records with missing and/or abnormal signal data, which can be caused by ECG system failure or external noise at the ECG collection, incorrect birth date or ECG collection time, or invalid weight and/or height.

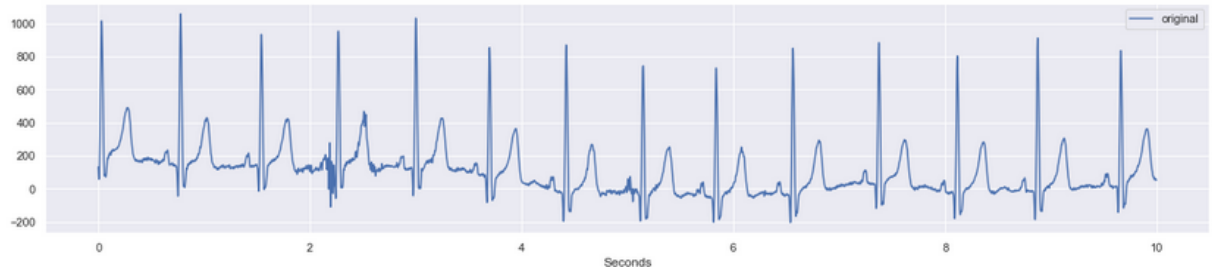


Figure 8.4 Example of a signal (the ECG lead) extract from a collected ECG

The signal extracted from the ECGs presented some noise, which is common in this kind of data acquisition, and even more usual when performed on children of early ages, due to motion and respiration movements. It was necessary to perform ECG filtering, before submitting data to ML models processing. Figure 8.4 presents an example of the extracted ECG signal.

Although feature engineering is usually a task performed in the data preparation phase, it was decided to create the derivative feature BMI (a derivative feature is a new feature which results from original features), as it was useful information to analyse at the current phase, including for anomalies and invalid data. BMI was calculated according to the standard formula (BMI = weight (kg) * height $(m)^{-2}$) and BMI-for-age values were classified according to the WHO reference data for BMI z-score into the following categories: normal weight (≤+1 SD, including only one child with thinness), overweight (>1 SD and ≤+2 SD) and obesity (>2 SD): The detail of classes is presented in Table 8-2. [280].

Table 8-2 BMI classification considering BMI measure

BMI classification		**Normal**	**Overweight**	**Obesity**
BMI score	Female	< 19	≥ 19 & < 22.6	≥ 22.6
	Male	< 18.5	≥ 18.5 & < 21.4	≥ 21.4

The histogram of BMI distribution, in males and females, and the bar chart for the BMI categories, in males and females, are shown in the Figure 8.5.

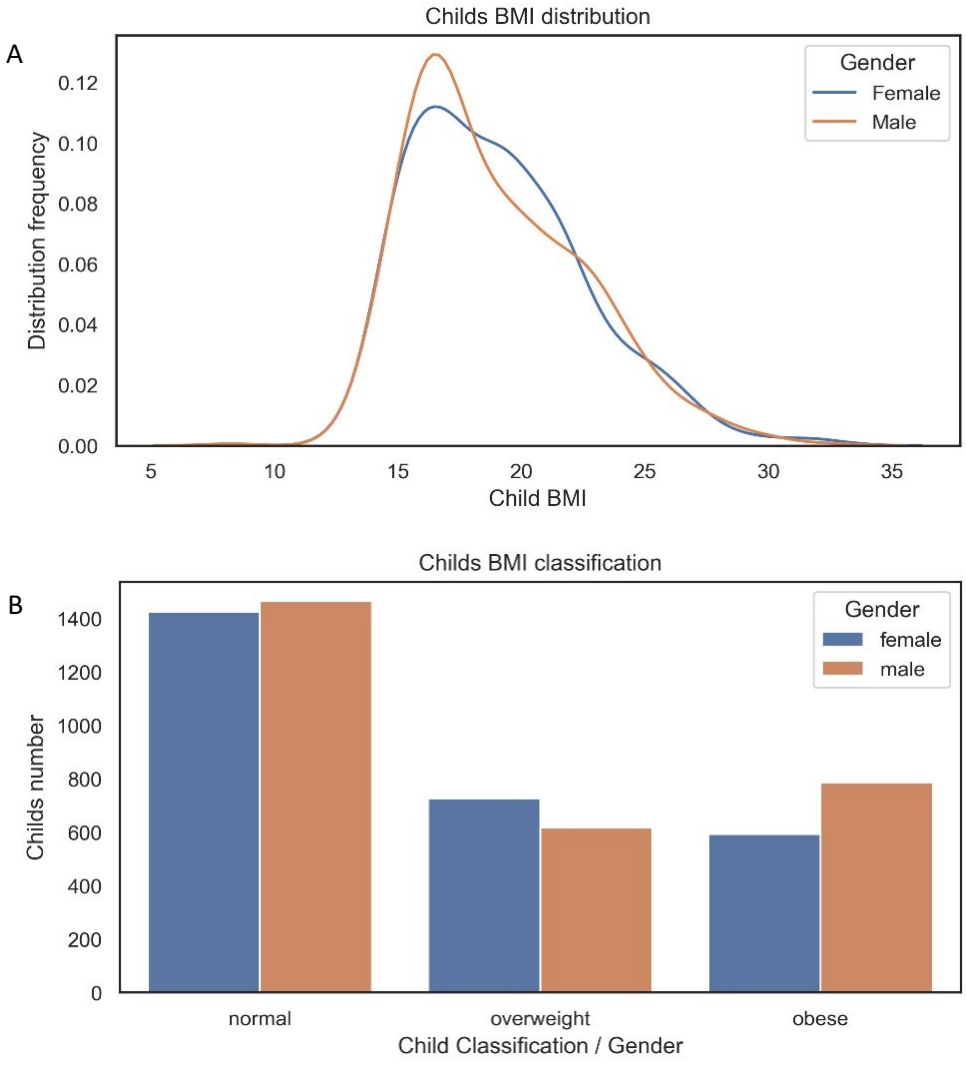


Figure 8.5 BMI statistical analysis (distribution and counting)

The Figure 8.5, plot A, highlights the BMI distribution with gender segmentation. We can observe that both curves, despite the gender, present a Gaussian or normal distribution, with some skew at right. The distribution shapes are quite similar.

The Figure 8.5, plot B, presents the number of individuals by BMI category and gender. In our sample, around 22/26% were categorized as overweight and 27/22% as obese, in males/females, respectively.

8.5 Data Preparation

The data preparation phase includes several tasks to reach the final dataset to be model, whose quality is crucial for the success of the models' outcomes.

The aim of this study was to determine if obesity was associated with ECG changes in the studied population. To reduce the noise among the data and to improve the study outcomes, it was decided to compare the obese group with the normal weight group. Records from overweight individuals were decided to be excluded from the analysis.

Each extracted ECG record was enriched by the medical report of the exam. The report was always performed by a paediatric cardiologist, in a text format, summarizing any alterations observed on the ECG. To focus on the study purpose, it was decided to remove any potential source of noise brought on by "external" factors. Those, all records that had been reported as abnormal were excluded from the final analysis. The most abnormalities found in the cohort's ECGs were arrhythmias, such as early repolarization patterns and bradycardia, auriculoventricular heart block, ventricular hypertrophy, or electrical axis deviation. Following this task, the team proceed to signal preparation.

The previous data understanding phase has identified the need to apply ECG filtering to records signal. The signal was processed with a finite impulse response (FIR) with a band-pass filter.

The resulting signal, with a 10.000 millisecond duration, was then decomposed on beats. For achieving the beats extraction, the first step was to identify R-peaks in the full signal. To this study, we have considered only “complete” beats, in which a complete beat is the range of signal composed by an R peak preceded by an interval of 100 samples, corresponding to 200ms and proceeded by 200 samples to which correspond 400ms. A total of 600ms, 300 samples where a “complete beat” corresponds to a PQRST complex. Every beat that did not fulfil these requirements, for example, because the R peak occurs near the record start or end, was discarded avoiding dealing with “incomplete” heart beats.

The team's analysis of the ECGs led to the range of 600ms, 200ms before R peak, and 400ms after peak. With this time frame, it is possible to ensure that the entire PQRST complex is captured and to maximise the extraction of valid dataset segments (heart beats) for the following phases.

Figure 8.6 presents an example of the signal processing tasks. The figure shows the ECG signal after FIR filtering and the identification of R-Peaks in “complete” beats.

A finite impulse response (FIR) filter in signal processing is one that settles to zero in a finite amount of time, making its impulse response (or reaction to any input of finite length) of finite duration. In contrast, infinite impulse response (IIR) filters have the potential to respond endlessly and may provide internal feedback, often fading [281].

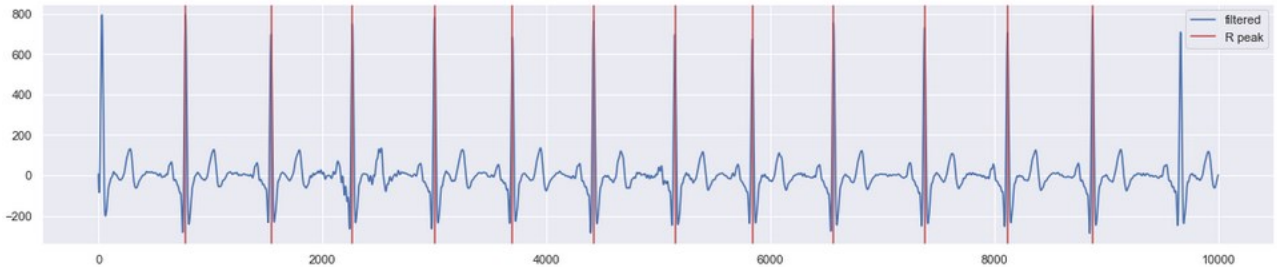


Figure 8.6 Example of ECG signal filtered with R-Peak and "complete" beats identification

The process resulted in the extraction of a total of 14.308 complete beats, with a 600ms duration (300 samples), which were considered valid according to the requirements defined before.

A last analysis was performed to enriched data understanding. The dataset beats signal was plotted on a graph (Figure 8.7), stratified by BMI classes (obese or normal), with the mean values and standard deviations bands being depicted.

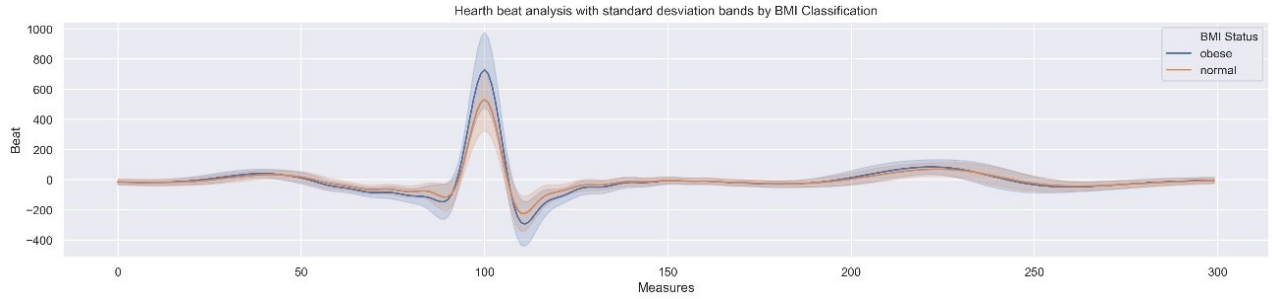


Figure 8.7 Beats signal mean and standard deviation bands, stratified by BMI classes.

The blue solid line represents the signal average for obese individuals while the orange line presents the average for normal weight children. The colour areas around the lines represent the standard deviation bands, which is limited by the standard deviation of the signal for each segment.

By analysis of Figure 8.7, some differences between obese and normal children beats seem to be evident. The chart with gender categorization was similar, with gender not affecting the distribution. A final dataset was created, including complete valid beats, according to the considered BMI classes.

8.6 Modelling

The study challenge encompasses the need for developing a model able to distinguish beats collected from normal or obese individuals. Considering the problem complexity in which the beat signal orientation/shape/trend is important, that is, the signal samples at each time/observation cannot be seen in isolation, it was decided to develop a model supported by neural networks (NN).

Considering the challenge characteristics, in particular the importance of observing samples not as independent elements and the team knowledge, convolutional neural networks (CNN) were pointed as the most promising ML architecture, namely with the application of 1D convolutions. Still, at this point, all possible NN architectures were considered.

1D CNN was deemed the most promising approach, since it allows deriving features from a fixed-length segment of data, in this case each of the heartbeats. The process allows learning from the signal data directly, without having to build explicit features.

The discovery of an optimized NN is an iterative, complex and time-consuming process, and thus it was decided to address it in detail in the following section.

In the ML modelling process, the best practices demand to split the final dataset into training, validation, and testing data. One of the most common approaches was used, dividing the dataset on 70/20/10, respectively.

8.7 Architecture Optimization

As previous mentioned, finding the best ML algorithm, meaning the optimized NN to fit the research data is a challenge by itself.

The NN algorithms family includes two types of hyperparameters, namely training and structural hyperparameters. The structure includes decisions about the NN number of layers (deep), number of units in each layer, kernel size, number of filters, stride, pooling, normalization techniques or activation functions. The training hyperparameters include decisions about optimizers, learning rate, learning rate momentum, batch size, initial weights, epochs, or early stopping patience. The existence of many variables (not all were mentioned) results in a nearly endless set of combinations.

For addressing the challenge of finding the best NN, namely the best classification accuracy, it was decided to perform an NN hyperparameters combination search, which is also known as Neural Architecture Search. The subject in best detailed in the chapter 4.1 Architecture Optimization .

The NAS process was performed on a machine with the following characteristics:

- OS: Ubuntu 18.04
- Disk: 500 GB SSD
- Sys Mem: 32Gb
- CPUs: AMD Ryzen 7 3600X
- GPUs: NVIDIA GTX 1080 TI 11GB + RTX 2060 SUPER 8GB

Several search approaches were applied to the process, namely search based on greedy, Bayesian, hyperband, random and genetic algorithms. NAS setting highlights: epochs: 200; max trials: 400; generation's number: 20; population size: 20. Search's goal: maximize validation accuracy.

The result of the NAS process applied to this challenge resulted in the following NN, in this case a CNN (Figure 8.8):

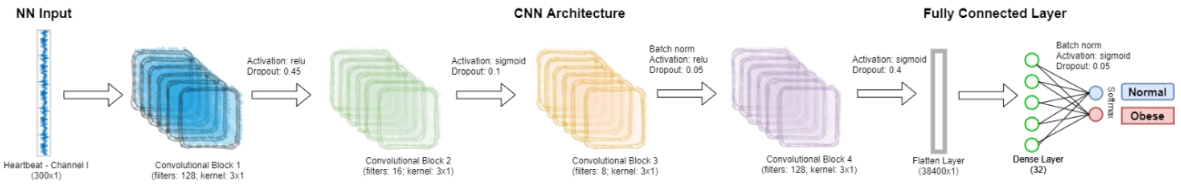


Figure 8.8 NAS process outcome, the best NN architecture for the research problem

The neural network structure is the following:

- Input layer: 300 inputs, corresponding to the length/size for every heartbeat (300 samples to which corresponds 600ms)
- Convolutional block 1:
    - Composed by 128 filters, 1-D kernel with size 3 and stride = 1, a ReLU activation function and a dropout of 0,45
- Convolutional block 2:
    - Composed by 16 filters, 1-D kernel with size 3 and stride = 1, a Sigmoid activation function and a dropout of 0,1
- Convolutional block 3:
    - Composed by 8 filters, 1-D kernel with size 3 and stride = 1, a batch normalization layer, a Relu activation function and a dropout of 0,05
- Convolutional block 4:
    - Composed by 128 filters, 1-D kernel with size 3 and stride = 1, a Sigmoid activation function and a dropout of 0,4
- Flatten layer: 38400 neurons
- Fully connected/dense block: 32 neurons
- Output layer: 2 classes (normal and obese), with a Softmax function

With the use of the mathematical function known as Softmax, a vector of numbers is transformed into a vector of probabilities, where the probability of each value is proportional to the relative scale of each value class. The probabilities in the vector sums to one for all possible outcomes or classes. This function is usually applied in classification problems.

The training hyperparameters were the following:

- Optimization function: AdamOptimizer
- Learning rate: 10−3
- Batch size: 200
- Early stopping criteria: 30 optimization steps based on validation set
- Loss function: categorical cross-entropy
- Weights initialization: Xavier initializer
- Bias initialization: zero

The achieved CNN presented the following performance, presented in Table 8-3 and Table 8-4:

Table 8-3 Confusion matrix performed on validation dataset

		Prediction	
		Normal	**Obese**
True Value	Normal	2018	99
	Obese	544	200

Table 8-4 Dataset's classification accuracy

	Training	**Validation**	**Testing**
Model Accuracy	78%	78%	79%

Table 8-3 presents the confusion matrix for the validation dataset. This matrix allows having a quick and complete view of the model outcomes produced for both classes, normal and obese. We can observe that among the predictions a total of 2218 records have been classified correctly (the sum of underline values) to which corresponds the validation accuracy of 78% (equation 8.1) presented on the following table. The accuracy achieved for the three sub-datasets, namely training, validation, and testing, is summarised in Table 8-4.

The achieved results were further evaluated on the following evaluation phase.

8.8 Evaluation

The previous modelling and optimization phases were able to achieve a model with very interesting performances, with accuracies of 78/79% - equation 8.1 - (highlight for the 79% at the testing dataset) on a problem that includes the use of complex data, namely signal representation based on real data, which always have the challenges of “real life” scenarios, namely errors, noise, and an “imperfect” research environment. From a technical perspective, the results can be considered very exciting.

Regarding the business goal, the model was able to classify the heartbeats with good accuracy, meaning that the beats signal provides relevant information to differentiate both classes. In another way, we can affirm that the normal and obese beats groups presented differences that can be “observed”/extracted by the model to classify them correctly.

The achieved results allow to address the research goals, revealing that some ECG differences exist and can be identified in obese individuals, when compared to the normal individuals group.

Still, when analysing the validation confusion matrix, in Table 8-3, it is possible to verify that among the obese records not all were correctly identified, with some being misclassified as normal. These results were discussed, considering both technical and clinical perspectives.

The current model considers the state of the individual at the ECG collection moment, and thus there is no information about the BMI status duration. Therefore, it was not considered for the present analysis the duration of obesity in the individuals classified as obese. The introduction of this information to the model would probably lead to an increase in the model performance and allow to extend the research results. That information was not available at the time the research was performed but could be considered in the future.

Until recently NN family algorithms were able to produce outcomes but would not provide additional details on how the outcomes were built. These models were considered “black boxes” and for obtaining more insights on how the results were produced, new models were required, resorting to different algorithms family (e.g.: decision trees). These algorithms are easier to understand but often entail a considerable performance decrease.

In this study, a different approach was taken (maintaining the model and the model performance benefits) which consisted of exploring the model through eXplainable Artificial Intelligence techniques. The application of XAI is presented in detail in the next section.

8.9 Explanation

In the past years, the concept of XAI has gained interest from the machine learning scientific community, as a way to obtain insights on how the outcomes of models are produced/generated. This interest grows with the complexity of models and the difficulty in understanding them, which is one of the challenges of deep learning.

XAI can be resumed as the application of methods that allow humans to understand and interpret the results/outputs generated by AI models. A machine learning model no longer has to be a "black-box" and model decisions can be explained.

A more formal definition for XAI could be "a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners". This subject is further detailed in 4.2 Explanation section [282] [283].

Considering the study characteristics, the goals to achieve, the model produced on the previous modelling phase (a CNN model), and the research made by the team, it was decided to explore the model using the XAI method Gradient-Weighted Class Activation Mapping Plus Plus (Grad-Cam++). Grad-Cam methods are generally applied to images, generating heatmaps which allow the visualization of areas with high importance on convolutional network architectures [284].

In a resumed way, the Grad-CAM method uses a weighted combination of the positive partial derivatives of the last convolutional layer feature maps, with respect to a specific class score, to generate a visual explanation for the class label under consideration. Chapter 4.2.9 Saliency Maps, Grad-Cam and Variations provides more detail information about this technique.

With the support of the technique the following histograms, presented in Figure 8.9 where generated. The images show the "key" signal moments for classifying the two classes of heartbeats: plot A for normal and plot B for obese. In other words, the most relevant inputs for the model to use in defining the classification of the record are the samples indicated by the histograms. The relevance is determined by the histogram bar's size.

To reduce noise and entropy on this analysis, the Grad-Cam++ technique was applied only to records that were correctly classified by the developed model.

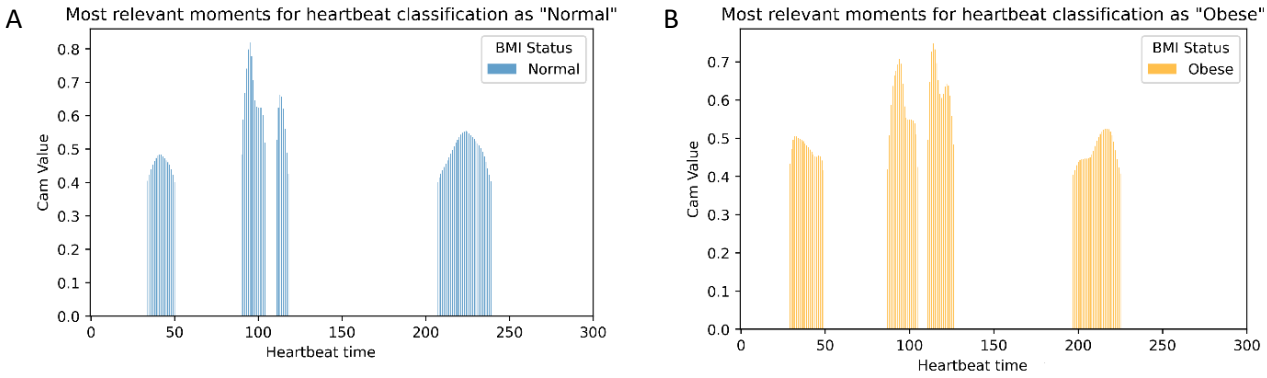


Figure 8.9 Most relevant moments for heartbeat classification using Grad-CAM++: A) Normal B) Obese

With a detailed observation, it is possible to verify zones of the two classes which overlap. From a technical perspective, that means that the model classifies the input "looking" at some areas that are common to both classes. That is, the model can find relevant differences between the two classes in those areas. The overlap areas and consequently the relevant distinctions between the two classes, which are reviewed and discussed in detail in the last framework phase (8.11 Research Results).

8.10 Deployment

The current study incorporates the goal of developing a machine learning model able to classify individuals (in this case children) heartbeats as obese or normal, though not with the purpose of making the model classification available for professionals, but instead as a base for research/analysis purposes.

Despite not being an initial goal of the project, would be possible to deploy the implemented model, in a real scenario, which would allow the analysis of children's heartbeats, to identify subtle changes in ECG that could be associated with the BMI status. The proposed model can be further developed, to serve as a tool to monitor ECG changes in children and adolescents, with obesity or other conditions.

The ML model creation has been developed as part of a data science techniques combination that allows obtaining information about the study research questions. Thus, this work will not result in an IT solution deployment, but instead in research results, which will be presented in the succeeding chapter.

8.11 Research Results

This research had the goal of understanding if some cardiovascular changes at the ECG records could be found in obese children, around the age of 10, and if these changes would be possible to identify, using data science techniques.

The evaluation chapter, complemented by the explanation chapter, were able to 1) confirm the existence of cardiovascular changes in obese children, in which a "smart" model was able to differentiate between obese and normal heartbeats with no other information besides the heartbeat signal; and 2) identify the heartbeat moments/zones which are more relevant for the model classification, and consequently those that present relevant alterations.

Figure 8.10, presented below, starts by joining the most relevant moments for the classification of each of the two classes, merging the information presented on Figure 8.9 (explanation section). This representation allows to perceive the overlapping between the two classes (upper plot). The bottom graph "captures" the classes overlap, presenting the signal regions which are relevant to both classes' classification. The overlap dimension was obtained from an average of both normal and obese classes Grad-CAM values.

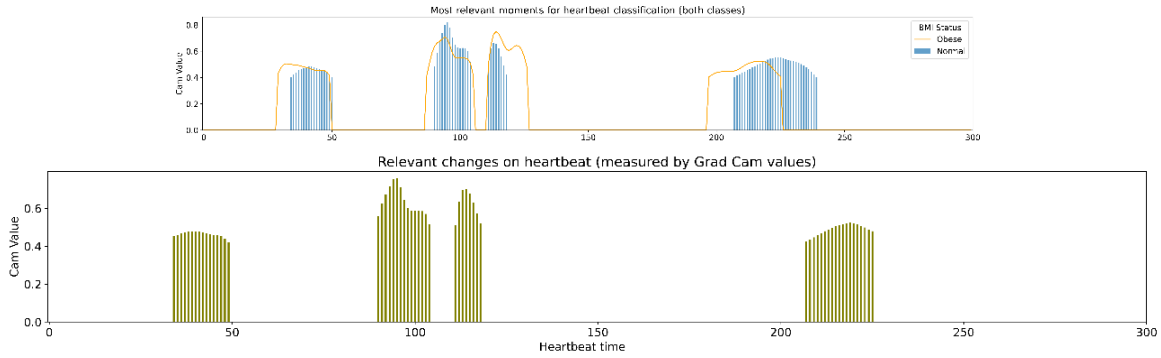


Figure 8.10 Grad-CAM analysis of both classes

The bottom graph presented in Figure 8.10 is the visual representation of the research goal, namely the identification of the regions where obese heartbeats present changes.

Taking as base both Figure 8.10 and Figure 8.7 (analysed still on the data preparation phase), we probably achieve the best representation of the research outcomes, which are summarized in Figure 8.11.

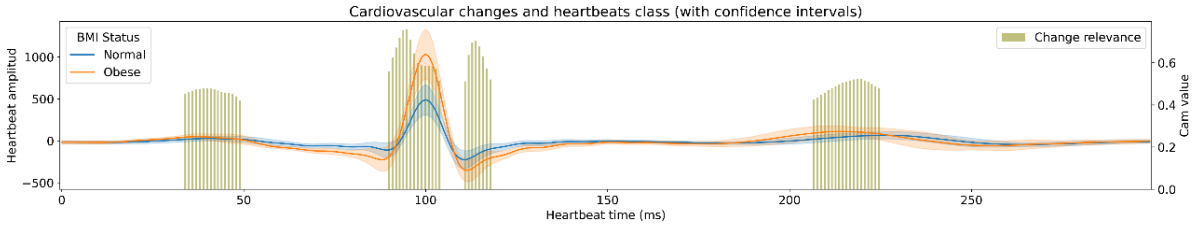


Figure 8.11 Grad-CAM results with the representation of the two heartbeat classes shapes

Figure 8.11 joins together the result of the Grad-CAM++ technique (performed on the explanation section), identifying the regions of the heartbeat signal which the machine learning model considers most relevant to classification, and consequently the ones that present the most significant differences, and the visual representation of the average and interval of confidence of the heartbeat signal for both classes.

With a detailed observation it is possible to match the signal moments with the typical heartbeat samples also known as PQRST complex. It is possible to understand that the model identifies cardiovascular changes in the regions between Q and R, immediately following S, and at the R peak. There are also less prominent alterations in the P wave and what appears to be a T wave anticipation in obese patients.

The figure also reveals several points where the classes signal crosses each other, which present a close relation with the XAI application outcomes, represented by the histogram bars.

All research findings were developed and validated by a multidisciplinary team which, besides the thesis team, also included a clinical team from the Generation XXI project.

8.12 Conclusions

The research presented emerges from a worldwide health challenge, the obesity pandemic and all associated consequences, with emphasis on cardiovascular comorbidities (e.g.: heart failure and coronary heart disease).

According to the literature review, there are several scientific studies on obesity, approaching diverse perspectives, such as aetiology, risk factors, lifestyle habits and genetic role. However, these studies usually emphasize on adulthood and respective medium to long-term consequences. Evidence supporting early cardiovascular changes due to obesity through paediatric age is scanty.

This study aimed to characterize possible cardiovascular changes due to obesity, in children at the age of 10-years old, by using data science techniques on analysing ECGs from a Portuguese population-based birth cohort. This research allowed to develop a machine learning classification to classify each heartbeat (collected from ECG) as from a normal weight or obese children and describe subtle cardiovascular changes on heartbeats. The combination of techniques permitted to verify the existence of cardiovascular abnormalities in obese children and identified the heartbeat areas/zones presenting changes.

The data science process has included SotA techniques to best explore data and extract valuable insights. Highlight for the use of NAS techniques on 1D data in a way to build the best possible machine learning model, able to fit the study data; and the applicability of XAI methods adapting/expanding existing techniques for 2D data to the study data, the signal on 1D shape.

The research was performed with a real dataset which incorporates an interesting population (random, diverse and with sufficient data to produce results), but more studies, if possible, applying the same proposed techniques, in other populations are welcome and can add more contributes to knowledge enrichment. For instance, the combination of the same techniques applied to a population of another country would make it possible to analyse eventual results similarities and/or differences.

The process here present was applied to a medical diagnostic method based on a signal analysis. The technique could also be used in other signal-based auxiliary/complementary exams, such as electroencephalography, electromyography or electrodermal activity. These are possible forms of evolution of the current application of techniques.

CHAPTER 9 - Conclusions & Future Work

9.1 Conclusions

This thesis's main research question was about the possibility to use pre-existing data, data science methodologies and tools, along with machine learning algorithms, to extract knowledge and value that can be used to improve the health and well-being of our society? Particularly, in this case emphasis on children was considered. The PhD research work developed during several years conducted to the conclusion that yes, it is possible.

The research was able to extract value and knowledge from already-existing data, demonstrating that it is possible to extract valuable and useful insights that were “hidden” in data until now. The PhD thesis presents real examples, as well as it provides resources for new initiatives in the health data science field. The potential to improve society's health and well-being is “now” (almost) limitless. Instead of creating new data collection processes, society can make use of the vast amount of already-existing, highly-potential data.

The thesis main contributions are expressed by:

- A new data science framework (an extension of an existing methodology). These contributions, which include the use of new methods that were already disseminated in proceedings and scientific journals with peer review, can be used in future data science research initiatives in the field of healthcare.
- Based on to the suggested framework, the thesis applied data science innovative techniques and methods developed three (successful) case studies related with children health care and well-being. All three cases are presented and discussed.
- The child's target height prediction study provided an improved model (when compared with SotA models) for the target height forecast, an important tool for assess the proper growth in children. Additionally, has provide a clearer knowledge of the role that genetics (parents’ characteristics) play in the process. The produced model may be applied to diverse populations, enabling paediatricians to monitor "target height" in a personalised manner. Specifically for Portugal, the thesis as engaged in a partnership with Instituto dos Registos e do Notariado, which has the custodian of numerous biometric data regarding the Portuguese population, in a way to produce a tailored model for the Portuguese population. Until now was not possible to deliver the customized model, within the timeframe of this thesis, due to the complexity involved in making the data available.

- The case study length of stay (LOS) in neonatal intensive care units (NICU) allowed identifying the factors influencing the length of stay in NICUs, as well as the extension of their impact. These two dimensions, relevance and impact, analyses represent a novelty in the LOS area, provided by the study approach.

  In addition to providing a better understanding of the variables determining LOS, the research offers information that can be vital for taking action to reduce LOS at NICUs and the associated effects and consequences. Future studies with similar characteristics (regarding LOS or in other health areas), may explore the suggested approach, avoiding the need to use traditional statistical inference/analysis projects, and producing faster results with more detailed information.
- The third use case was focused on potential cardiovascular changes/abnormalities in children brought by obesity. Two unanswered clinically questions regarding the health and wellbeing of children were addressed. The study was able to uncover the existence of effects due childhood obesity and identify their cardiovascular impacts. The introduced techniques can be used for both children and other demographic groups, in different health scenarios that may be tracked/monitored by signal data (1D data).
- Several architecture optimization procedures were developed, including NAS processes for 1D CNN (e.g: NAS for 1D CNNS with genetic algorithms, or optimization for LightGBM). The advancements were shared with the community, allowing other researchers to use the resources in other NAS and optimization challenges, including for fields unrelated to health.
- Beside the new developments, it was also performed a benchmark which compares the different strategies of 1D CNN NAS. A complete analysis was provided to the scientific community with the publication of a paper in a relevant scientific peer-to-peer journal.
- The Grad-Cam technique, usually applied in 2D (image) was adapted by the thesis to 1D data, allowing to provide explainability also in this type of data.
- Several scientific papers were produced. Five of them have already been published in peer-reviewed journals, and two more are being prepared to be submitted. Highlight for the fact that two of the research works were published in journals ranked in the first quarter.

9.2 Future Work

The process of data science makes it fundamentally dependent on data. The studies developed in the PhD thesis context were based on randomly selected populations and the collection processes were not data science oriented, but rather oriented for other kind of processes (typically clinical processes, but definitely not with a data science research purpose).

Reusing these types of datasets is considered an add value because they represent an interesting segment of the (paediatric) population without being biased by any pathology or disease. This may contribute to an ML model that has a greater capacity to generalise the results that are produced. However, because real datasets have issues like incomplete, irregular, incoherent, and inaccurate data, among other problems, this type of datasets also represents an extra challenge.

It is important to emphasise that the thesis research was always developed under the adoption of data science best practices, which mitigate data flaws and contribute to the generalizability of the results.

Nevertheless, it is crucial that in the future, the ideas developed in the performed studies may be applied to other populations, either as a study or in practise, and that the findings will be analysed by comparison to the actual scenarios.

In the field of data science, the ability to customise and specialise models comes with a "cost" that may lead to the necessity to adjust the models that have been created. This may be required, for example, to ensure improved compliance with the new scenarios, to adapt to new researches characteristics, or to deploy models into practise. For instance, the population in a new study may have different characteristics that cause variations in the models. In that case, the need for readjustment to a new population or scenario, the thesis still provides the procedures and resources needed to make the required changes.

For instance, considering the use case that analyses cardiovascular changes due to obesity in children's. The research, which involved 10-year-old children, produced scientific conclusions, but new questions quickly emerged. Can the cardiac consequences of obesity be reversed with treatment? What are the consequences at different ages, like 5, 15, and 20 years old? It would be interesting to conduct an update of the study addressing children at different ages or, if it's feasible, to repeat the study with the same population at different ages.

Combining the processing of the ECG signal (1D data) with the processing of the same data, but in image format (2D data) could be another research opportunity. The aim of this exercise would be to evaluate if it is possible to extract different/more information from the same data, but observing it in different formats (signal and image), benefiting from the development of deep learning in the image analysis area.

Another suggestion from this thesis is to apply the proposed framework and techniques to other diagnostic tools that, such as the electrocardiogram (ECG), are based on 1D physiological signals. Examples include Electrodermal Activity (EDA), Photoplethysmogram (PPG), Electroencephalography (EEG), and Electromyography (EMG).

Regarding the hospitalization days in Neonatal Intensive Care Units, the presented results are directly applicable at the national level (Portugal) and their generalisation has been validated. It would be of great interest to re-apply the study on data from another country (or region of the country with significant size), assessing whether the causes and consequences of the length of stay at neonatal hospital are the same across countries, or whether there are specificities and their origin. Once again, this thesis provides the resources needed to develop this work.

Nevertheless we consider that the thesis' goals have all been successfully met, we still face the challenge of exposing and disseminating these concepts and findings to the scientific community, as well as to the medical community. In this data science field, collaboration between experts is essential, and the only way to put the concepts that have been introduced into practise is to establish partnerships involving data science teams and physicians.

In addition to applying the proposed framework and the techniques and processes developed to other health-related scenarios and problems, demonstrating their capacity to produce value and knowledge, it will be valuable to continue progressing in the field with new data science approaches, methods and techniques for healthcare and well-being.

In this last point, we stress the XAI and the architecture optimization, where the NAS concept is included. Although the thesis presents developments and applications in these fields, it is clear that these recent concepts offer an enormous potential, and research will undoubtedly lead to new discoveries in a near future. Hopefully, after this thesis, the team will continue to contribute to this scientific progress.

References

[1] J. R. Cordeiro, O. Postolache and J. C. Ferreira, “"Child’s Target Height Prediction Evolution,” *Applied Sciences ,* vol. 9, no. 24, p. 5447, 2019.

[2] J. R. Cordeiro and O. Postolache, “Length of Stay Analysis at Neonatal Care Units with Data Science - Preliminary Results,” *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA),* pp. 1-6, 2021.

[3] J. R. Cordeiro, A. Raimundo, O. Postolache and P. Sebastião, “Neural Architecture Search for 1D CNNs—Different Approaches Tests and Measurements,” *Sensors,* vol. 21, p. 7990, 2021.

[4] J. R. Cordeiro and O. Postolache, “Big Data Storage for a Health Predictive System,” *2018 International Symposium in Sensing and Instrumentation in IoT Era (ISSI),* pp. 1-6, 2018.

[5] O. Postolache, L. Teixeira, J. R. Cordeiro, L. Lima, P. Arriaga, M. Rodrigues and P. Girão.

[6] J. Mcginnis, P. Russo and J. Knickman, “The Case For More Active Policy Attention To Health Promotion,” *Health affairs,* vol. 21, pp. 78-93, 03 2002.

[7] J. Fradinho, D. Nightingale and M. & Fradinho, “A Systems-of-Systems Perspective on Healthcare: Insights From Two Multi-Method Exploratory Cases of Leading U.S. and U.K. Hospitals,” *IEEE Systems Journal,* vol. 8, no. 3, pp. 795-802, 09 2014.

[8] Y.-K. Chen and Y.-H. Tao, “Analyzing the Healthcare Expenditure of National Health Insurance: A HLM Model Construction on the National Health Insurance Database,” *12th International Conference on Advances in Mobile Computing and Multimedia, MoMM 2014,* pp. 353-355, 12 2014.

[9] C. Van Hoof and J. Penders, “Addressing the healthcare cost dilemma by managing health instead of managing illness - An opportunity for wearable wireless sensors,” *Proceedings - Design, Automation and Test in Europe, DATE,* pp. 1537-1539, 2013.

[10] D. Di Pasquale, M. Padula, P. L. Scala, L. Biocca and N. Paraciani, “Advancements in ICT for healthcare and wellbeing: towards Horizon 2020,” *Proceedings of the 5th International Conference on Management of Emergent Digital EcoSystems, MEDES 2013,* pp. 353-358, 10 2013.

[11] R. L. Jessup, D. A. O'Connor, P. Putrik, K. Rischin, J. Nezon, S. Cyril, S. Shepperd and R. Buchbinder, “Alternative service models for delivery of healthcare services in high-income countries: a scoping review of systematic reviews,” *British Medical Journal Publishing Group,* vol. 9, no. 1, 2019.

[12] C.-G. S., A.-L. M. and A.-L. B., “The Demand for Healthcare Services and Resources: Patterns, Trends and Challenges in Healthcare Deliver,” *Springer Proceedings in Mathematics & Statistics,* vol. 278, 2019.

[13] World Health Organization, “COVID-19: operational guidance for maintaining essential health services during an outbreak: interim guidance,” *World Health Organization.*

[14] H. Atasoy, B. N. Greenwood and J. S. McCullough, “The Digitization of Patient Care: A Review of the Effects of Electronic Health Records on Health Care Quality and Utilization,” *Annual review of public health,* vol. 40, p. 487–500, 2019.

[15] E. Kim, S. M. Rubinstein, K. T. Nead, A. P. Wojcieszynski, P. E. Gabriel and J. L. Warner, “The Evolving Use of Electronic Health Records (EHR) for Research,” *Seminars in Radiation Oncology,* vol. 29, no. 4, pp. 354-361, 2019.

[16] F. Girardi, G. De Gennaro, L. Colizzi and N. Convertini, “Improving the Healthcare Effectiveness: The Possible Role of EHR, IoMT and Blockchain,” *Electronics,* vol. 9, no. 6, 2020.

[17] A. K. Yetisen, J. L. Martinez-Hurtado, B. Ünal, A. Khademhosseini and H. Butt, “Wearables in Medicine,” *Advanced Materials,* vol. 30, no. 33, 2018.

[18] N. Niknejad, W. B. Ismail, A. Mardani, H. Liao and I. Ghani, “A comprehensive overview of smart wearables: The state of the art literature, recent advances, and future challenges,” *Engineering Applications of Artificial Intelligence,* vol. 90, 2020.

[19] J. Wang, Y. Yang, T. Wang, R. S. Sherratt and J. Zhang, “Big Data Service Architecture: A Survey,” *Journal of Internet Technology,* vol. 21, no. 2, pp. 393-405, 2020.

[20] R. Hariri, E. Fredericks and K. Bowers, “Uncertainty in big data analytics: survey, opportunities, and challenges,” *Journal of Big Data,* vol. 6, no. 44, 2019.

[21] R. Wason, “Deep learning: Evolution and expansion,” *Cognitive Systems Research,* vol. 52, pp. 701-708, 2018.

[22] W. Wang and S. Keng, “Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda,” *Journal of Database Management (JDM),* vol. 30, no. 1, pp. 61-79, 2019.

[23] A. Ng, AI is the new electricity, O'Reilly Media, 2018.

[24] D. Gündüz, P. de Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy and M. van der Schaar, “Machine Learning in the Air,” *IEEE Journal on Selected Areas in Communications,* vol. 37, no. 10, pp. 2184-2199, 2019.

[25] S. J. Qin and L. H. Chiang, “Advances and opportunities in machine learning for process data analytics,” *Computers & Chemical Engineering,* vol. 126, pp. 465-473, 2019.

[26] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Computer Science,* vol. 160, 2021.

[27] B. Marr and M. Ward, Artificial intelligence in practice : how 50 successful companies used AI and machine learning to solve problems, Chichester, 2019.

[28] “Machine Learning and Deep Learning Applications-A Vision,” *Global Transitions Proceedings,* vol. 2, no. 1, pp. 24-28, 2021.

[29] D. Tsay and C. Patterson, “From Machine Learning to Artificial Intelligence Applications in Cardiac Care,” *Circulation,* vol. 138, no. 22, pp. 2569-2575, 2018.

[30] S. Lalmuanawma, J. Hussain and L. Chhakchhuak, “Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review,” *Chaos, Solitons & Fractals},* vol. 139, 2020.

[31] R. Manne and S. Kantheti, “Application of Artificial Intelligence in Healthcare: Chances and Challenges,” *Journal of Applied Science and Technology},* vol. 40, pp. 78-89, 2021.

[32] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman, S. Michiels, K. Souris, E. Sterpin and J. A. Lee, “Artificial intelligence and machine learning for medical imaging: A technology review,” *Physica Medica,* vol. 83, pp. 242-256, 2021.

[33] Kun-Hsing, A. L. Beam and I. S. Kohane, “Artificial intelligence in healthcare,” *Nature biomedical engineering,* vol. 2, no. 10, pp. 719-731, 2018.

[34] M. A. Alloghani, D. A.-J. Obe, J. Mustafina and A. Hussain, Supervised and Unsupervised Learning for Data Science, 2020.

[35] J. McCarthy, “Artificial Intelligence, Logic and Formalizing Common Sense,” in *Philosophical Logic and Artificial Intelligence*, Springer Netherlands, 1989.

[36] Saurabh Jha and E. J. Topol, “Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists,” *Journal of the American Medical Association,* vol. 316, no. 12, p. 2353–2354, 2016.

[37] K.-H. Yu and D. A. e. a. Levine, “Predicting Ovarian Cancer Patients’ Clinical Response to Platinum-Based Chemotherapy by Their Tumor Proteomic Signatures,” *Journal of Proteome Research,* vol. 15, no. 8, pp. 1535-3893, 2016.

[38] S. Secinaro, Davide-Calandr, A. Secinaro, V. Muthurangu and P. Biancone , “The role of artificial intelligence in healthcare: a structured literature review,” *BMC Medical Informatics and Decision Making,* vol. 21, no. 1, 2021.

[39] H. Z. Z. S. G. J. Y. F. Guo Y, “Artificial Intelligence in Health Care: Bibliometric Analysis,” *Journal of medical Internet research,* vol. 22, no. 7, p. e18228, 2020.

[40] R. A. Miller, “Medical Diagnostic Decision Support Systems—Past, Present, And Future: A Threaded Bibliography and Brief Commentary,” *Journal of the American Medical Informatics Association,* vol. 1, no. 1, p. 8–27, 1994.

[41] R. S. P. W. B. S. Peter Szolovits, “Artificial Intelligence in Medical Diagnosis,” *Annals of Internal Medicine,* vol. 108, no. 1, pp. 80-87, 1988.

[42] K.-H. Yu and M. Snyder, “Omics Profiling in Precision Oncology,” *Molecular & cellular proteomics,* vol. 15, no. 8, pp. 2525-36, 2016.

[43] D. S. Rigel, R. J. Friedman, . W. Kopf and . Polsky, “ABCDE - an evolving concept in the early detection of melanoma,” *Archives of dermatology,* vol. 141, no. 8, p. 1032–1034, 2005.

[44] A. L. Beam and I. S. Kohane, “Big Data and Machine Learning in Health Care,” *JAMA,* vol. 319, no. 13, p. 1317–1318, 2018.

[45] B. Mahesh, “Machine learning algorithms-a review,” *International Journal of Science and Research (IJSR),* vol. 9, pp. 381-386, 2020.

[46] C. Yu, J. Liu, S. Nemati and G. Yin, “Reinforcement Learning in Healthcare: A Survey,” *ACM Computing Surveys,* vol. 55, no. 1, 2023.

[47] K.-H. Y. e. al., “Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma,” *Cell systems,* vol. 5, no. 6, 2017.

[48] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin and e. al., “Opportunities and obstacles for deep learning in biology and medicine,” *Journal of the Royal Society Interface,* vol. 12, no. 141, 2018.

[49] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega and D. R. Webster, “Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,” *JAMA,* vol. 316, no. 22, p. 2402–2410, 2016.

[50] K. Zhou, W. Li and D. Zhao, “Deep learning-based breast region extraction of mammographic images combining pre-processing methods and semantic segmentation supported by Deeplab v3,” *Technology and health care : official journal of the European Society for Engineering and Medicin,* vol. 30, p. 173–190, 2022.

[51] N. Codella, Q.-B. Nguyen, S. Pankanti, D. Gutman, B. Helba, A. Halpern and J. Smith, “Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images,” *Ibm Journal of Research and Development,* vol. 61, no. 4, 2017.

[52] B. Cho, Y. Choi and M. e. a. Lee, “Classification of cervical neoplasms on colposcopic photography using deep learning,” *Scientific Reports,* vol. 10, no. 1, 2020.

[53] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura and R. M. Summers, “Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning,” *IEEE transactions on medical imaging,* vol. 35, no. 5, p. 1285–1298, 2016.

[54] A. Rajkomar, S. Lingam, A. G. Taylor, M. Blum and J. Mongan, “High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks,” *Journal of digital imaging,* vol. 30, no. 1, p. 95–101, 2017.

[55] P. Lakhani and B. Sundaram, “Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks,” *Radiology,* vol. 284, no. 2, p. 574–582, 2017.

[56] G. Amit, R. Ben-Ari, O. Hadad, E. Monovich, N. Granot and S. Y. Hashou, "Classification of breast MRI lesions using small-size training sets: comparison of deep learning approaches," *Medical Imaging,* 2017.

[57] B. v. Ginneken, A. A. A. Setio, C. Jacobs and F. Ciompi, "Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans," *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI),* pp. 286-289, 2015.

[58] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira and M. A. G. Lopez, "Convolutional neural networks for mammography mass lesion classification," *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* pp. 797-800, 2015.

[59] S. Zhu, . Gilbert, . Chetty and . Siddiqui, "The 2021 landscape of FDA-approved artificial intelligence/machine learning-enabled medical devices: An analysis of the characteristics and intended use," *International journal of medical informatics,* vol. 165, p. 104828, 2022.

[60] A. Choudhury and O. Asan, "Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review," *MIR medical informatics,* vol. 8, no. 7, p. e18599, 2020.

[61] R. Leaman and Z. Lu, "TaggerOne: joint named entity recognition and normalization with semi-Markov Models," *Bioinformatics (Oxford, England),* vol. 32, no. 18, p. 2839–2846, 2016.

[62] S. Liu, B. Tang, Q. Chen and X. Wang, "Effects of Semantic Features on Machine Learning-Based Drug Name Recognition Systems: Word Embeddings vs. Manually Constructed Dictionaries," *Information,* vol. 6, no. 4, pp. 848-865, 2015.

[63] "Extracting drug–drug interactions from no-blinding texts using key semantic sentences and GHM loss," *Journal of Biomedical Informatics,* vol. 135, p. 104192, 2022.

[64] Y. Peng, C.-H. Wei and Z. Lu, "Improving chemical disease relation extraction with rich features and weakly labeled data," *Journal of Cheminformatics,* vol. 8, no. 53, 2016.

[65] F. Li, Y. Zhang, M. Zhang and D. Ji, "Joint models for extracting adverse drug events from biomedical text," *IJCAI'16: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence,* p. 2838–2844, 2016.

[66] X. Yang, Y. Wang, R. Byrne, G. Schneider and S. Yang, "Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery," *Chemical Reviews,* vol. 119, no. 18, p. 10520–10594, 2019.

[67] S. Wang, . Sun, . Li, . Zhang and . Xu, "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model," *PLoS computational biology,* vol. 13, no. 1, p. e1005324, 2017.

[68] G. P. Moustris, S. C. Hiridis, K. M. Deliparaschos and K. M. Konstantinidis, "Evolution of autonomous and semi-autonomous robotic surgical systems: a review of the literature.," *The international journal of medical robotics + computer assisted surgery : MRCAS,* vol. 7, no. 4, p. 375–392, 2011.

[69] R. Elek, T. D. Nagy, D. Nagy, G. Kronreif, I. Rudas and T. Haidegger, "Recent trends in automating robotic surgery," *2016 IEEE 20th Jubilee International Conference on Intelligent Engineering Systems (INES),* 2016.

[70] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger and P. C. W. Kim, “Supervised autonomous robotic soft tissue surgery,” *Science Translational Medicine,* vol. 8, no. 337, pp. 337ra64-337ra64, 2016.

[71] T. M. Connelly, Z. Malik, R. Sehgal, G. Byrnes, J. C. Coffey and C. Peirce , “The 100 most influential manuscripts in robotic surgery: a bibliometric analysis,” *Journal of Robotic Surgery volume,* vol. 14, p. 155–165, 2020.

[72] P. Yadav, M. Steinbach, V. Kumar and G. Simon, “Mining Electronic Health Records (EHRs): A Survey,” *Association for Computing Machinery,* vol. 50, no. 6, p. 1–40, 2018.

[73] S. V. G. Subrahmanya, D. K. Shetty, V. Patil, B. M. Z. Hameed, R. Paul, K. Smriti, N. Naik and B. K. Somani , “The role of data science in healthcare advancements: applications, benefits, and future prospects,” *Irish Journal of Medical Science ,* vol. 191, no. 4, p. 1473–1483, 2022.

[74] X. Cai, O. Perez-Concha, E. Coiera, F. Martin-Sanchez, R. Day, D. Roffe and B. Gallego, “Real-time prediction of mortality, readmission, and length of stay using electronic health record data,” *Journal of the American Medical Informatics Association,* vol. 23, no. 3, p. 553–561, 2016.

[75] S. Majumder, T. Mondal and M. J. Deen., “Wearable Sensors for Remote Health Monitoring,” *Sensors,* vol. 17, no. 1, p. 130, 2017.

[76] E. Svennberg, F. Tjong, A. Goette, N. Akoum, L. Di Biase, P. Bordachar, G. Boriani, H. Burri, G. Conte, J. C. Deharo, T. Deneke, I. Drossart and D. Duncker, “How to use digital devices to detect and manage arrhythmias: an EHRA practical guide,” *EP Europace,* vol. 24, no. 6, p. 979–1005, 2022.

[77] M. Pastorino, M. T. Arredondo, J. Cancela and S. Guillen, “Wearable sensor network for health monitoring: the case of Parkinson disease,” *Journal of Physics: Conference Series,* vol. 450, no. 1, pp. 1742-6596, 2013.

[78] J. Takacs, C. L. Pollock, J. R. Guenther, M. Bahar and C. Napier, “Validation of the Fitbit One activity monitor device during treadmill walking,” *Journal of science and medicine in sport,* vol. 17, no. 5, p. 496–500, 2014.

[79] R. Yang, E. Shin, M. W. Newman and M. S. Ackerman, “When fitness trackers don't 'fit': end-user difficulties in the assessment of personal tracking device accuracy,” *2015 ACM International Joint Conference,* pp. 623-634, 2015.

[80] J. Yin, K. Y. Ngiam and H. H. Teo, “Role of Artificial Intelligence Applications in Real-Life Clinical Practice: Systematic Review,” *Journal of Medical Internet Research,* vol. 23, no. 4, p. e25759, 2021.

[81] R. C. Deo, “Machine Learning in Medicine,” *Circulation,* vol. 132, no. 20, pp. 1920-1930, 2015.

[82] M. T. Ribeiro, S. Singh and C. Guestrin, “Why Should I Trust You?": Explaining the Predictions of Any Classifier,” *arXiv,* 2016.

[83] V. Dignum, “Ethics in artificial intelligence: introduction to the special issue,” *Ethics and Information Technology,* vol. 20, p. pages 1–3, 2018.

[84] W. N. Price, “Artificial Intelligence in Health Care: Applications and Legal Implications,” *University of Michigan Law School,* 2017.

[85] B. Meskó, Z. Drobni, É. Bényei, B. Gergely and Z. Győrffy, “Digital health is a cultural transformation of traditional healthcare,” *Digital health is a cultural transformation of traditional healthcare,* vol. 3, no. 9, 2017.

[86] G. 2. C. a. A. H. Collaborators, “Diseases, Injuries, and Risk Factors in Child and Adolescent Health, 1990 to 2017: Findings From the Global Burden of Diseases, Injuries, and Risk Factors 2017 Study,” *JAMA Pediatr,* vol. 173, no. 6, 2019.

[87] B. TF and W. JA, “How Can the Pediatric Community Enhance Funding for Child Health Research?,” *JAMA Pediatr.,* vol. 175, no. 12, p. 1212–1214, 2021.

[88] Z. Vesoulis, A. Husain and F. Cole, “Improving child health through Big Data and data science,” *Pediatric Research,* 2022.

[89] J. Reyes-González, R. Díaz-Peregrino, S.-U. Victor, G.-R. Isabel, P. Castillo and E. Ogando-Rivas, “Big data in the healthcare system: a synergy with artificial intelligence and blockchain technology,” *Journal of Integrative Bioinformatics,* vol. 19, no. 1, 2021.

[90] C. Kelly, A. Brown and J. Taylor, “Artificial Intelligence in Pediatrics,” in *Artificial Intelligence in Medicine*, Springer International Publishing, 2021, pp. 1-18.

[91] K. Allegaert, S. H. P. Simons, D. Tibboel, E. H. Krekels, C. A. Knibbe and J. N. van den Anker, “Non-maturational covariates for dynamic systems pharmacology models in neonates, infants, and children: Filling the gaps beyond developmental pharmacology,” *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences,* vol. 109S, p. S27–S31, 2017.

[92] Z. T. A. L. H. e. a. Vesoulis, “Racial discrepancy in pulse oximeter accuracy in preterm infants,” *Journal of Perinatology volume,* vol. 42, p. 79–85, 2022.

[93] M. J. Taylor, E. S. Dove, G. Laurie and D. Townend, “When can the Child Speak for Herself? The Limits of Parental Consent in Data Protection Law for Health Research,” *Medical Law Review,* vol. 26, no. 3, pp. 369-391, 2017.

[94] J. M. Brussee, E. A. M. Calvier, E. H. J. Krekels, P. A. J. Välitalo, D. Tibboel, K. Allegaert and C. A. J. Knibbe, “Children in clinical trials: towards evidence-based pediatric pharmacotherapy using pharmacokinetic-pharmacodynamic modeling,” *Expert Review of Clinical Pharmacology,* vol. 9, no. 9, pp. 1235-1244, 2016.

[95] C. I. S. Barker, J. F. Standing, L. E. Kelly, L. Hanly Faught, A. C. Needham, M. J. Rieder, S. N. de Wildt and M. Offringa, “Pharmacokinetic studies in children: recommendations for practice and research,” *Archives of disease in childhood,,* vol. 103, no. 7, p. 695–702, 2018.

[96] S. Goulooze, L. Zwep, J. Vogt, E. Krekels, T. Hankemeier, J. van den Anker and C. Knibbe, “Beyond the Randomized Clinical Trial: Innovative Data Science to Close the Pediatric Evidence Gap,” *Clinical Pharmacology & Therapeutics,* vol. 107, pp. 786-795, 2020.

[97] K. B. Cohen, B. Glass, H. M. Greiner, K. Holland-Bouley, S. Standridge, R. Arya, R. Faist, D. Morita, F. Mangano, B. Connolly, T. Glauser and J. Pestian, "Methodological Issues in Predicting Pediatric Epilepsy Surgery Candidates Through Natural Language Processing and Machine Learning.," *Biomedical informatics insights,* vol. 8, p. 11–18, 2016.

[98] J. P. Pestian, M. Sorter, B. Connolly, B. Cohen, M. C. K. and e. al., "Machine Learning Approach to Identifying the Thought Markers of Suicidal Subjects: A Prospective Multicenter Trial," *Suicide & life-threatening behavior,* vol. 47, no. 1, p. 112–121, 2017.

[99] H. Liang, B. Tsui and H. e. a. Ni, "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence.," *Nature Medicine,* vol. 25, no. 3, 2019.

[100] G. Savova, J. Pestian, B. Connolly, T. Miller, Y. Ni and J. Dexheimer, "Natural Language Processing: Applications in Pediatric Research.," in *Pediatric Biomedical Informatics. Translational Bioinformatics*, Springer, 2016.

[101] H.-G. Eichler, B. Bloechl-Daum, K. Broich and e. al., "Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals?," *Clin. Pharmacol. Ther.,* no. 105, pp. 912-922, 2019.

[102] S. Shelmerdine, K. Rosendahl and O. Arthurs, "Artificial intelligence in paediatric radiology: international survey of health care professionals' opinions," *Pediatric Radiology volume,* vol. 52, p. 30–41, 2022.

[103] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage,* vol. 108, pp. 214-224, 2015.

[104] D. B. Larson, M. C. Chen, M. P. Lungren, S. S. Halabi, N. V. Stence and C. P. Langlotz, "Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs," *Radiology,* vol. 287, no. 1, pp. 313-322, 2018.

[105] R. M. Summers, "Deep Learning Lends a Hand to Pediatric Radiology," *Radiology,* vol. 287, no. 1, pp. 323-325, 2018.

[106] "Identifying facial phenotypes of genetic disorders using deep learning," *Nature Medicine,* vol. 25, no. 1, 2019.

[107] A. Nagori, L. S. Dhingra, A. Bhatnagar, R. Lodha and T. Sethi, "Predicting Hemodynamic Shock from Thermal Images using Machine Learning," *Scientific Reports volume,* vol. 9, no. 91, 2019.

[108] Z. Hoodbhoy, S. Masroor Jeelani, A. Aziz, M. I. Habib and e. al., "Machine Learning for Child and Adolescent Health: A Systematic Review," *Pediatrics,* vol. 147, no. 1, 2021.

[109] C. P. Travers, W. A. Carlo and N. Ambalavanan, "The Future of Outcome Prediction for Preterm Infants in the Neonatal ICU," *American journal of respiratory and critical care medicine,* vol. 205, no. 1, p. 6–8, 2022.

[110] A. Sheikhtaheri, M. R. Zarkesh, R. Moradi and F. Kermani, “Prediction of neonatal deaths in NICUs: development and validation of machine learning models,” *BMC Medical Informatics and Decision Making volume,* vol. 21, no. 131, 2021.

[111] “Neonatal Jaundice Detection System,” *Journal of Medical Systems,* vol. 40, no. 166, 2016.

[112] H. Jeong and R. Kamaleswaran, “Pivotal challenges in artificial intelligence and machine learning applications for neonatal care,” *Seminars in Fetal and Neonatal Medicine,* vol. 27, no. 5, p. 101393, 2022.

[113] A. Bertini, R. Salas, S. Chabert, L. Sobrevia and F. Pardo, “Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review,” *Frontiers in Bioengineering and Biotechnology,* vol. 9, 2022.

[114] M. Rahman, O. Usman, R. Muniyandi, S. Sahran, S. Mohamed and R. Razak, “A Review of Machine Learning Methods of Feature Selection and Classification for Autism Spectrum Disorder,” *Brain Sciences,* vol. 10, no. 12, p. 949, 2020.

[115] C. Costescu, B. Vanderborght and D. David, “Reversal Learning Task in Children with Autism Spectrum Disorder: A Robot-Based Approach,” *Journal of Autism and Developmental Disorders,* vol. 45, p. 3715–3725, 2015.

[116] J. Zhang, “Multivariate Analysis and Machine Learning in Cerebral Palsy Research,” *Frontiers in Neurology,* vol. 8, 2017.

[117] J. Jamnik, “Artificial intelligence in paediatrics: A checkup,” *University of Toronto Medical Journa,* vol. 96, 2019.

[118] IBM, “Introduction to CRISP-DM,” [Online]. Available: https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=guide-introduction-crisp-dm. [Accessed 2022/03/30].

[119] R. Wirth and J. Hipp, “CRISP-DM: Towards a standard process model for data mining,” *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining,* 2000.

[120] S. Huber, H. Wiemer, D. Schneider and S. Ihlenfeldt, “DMME: Data Mining Methodology for Engineering Applications-A Holistic Extension to the CRISP-DM Model,” *12th CIRP Conference on Intelligent Computation in Manufacturing Engineering,* vol. 79, pp. 403-408, 2019.

[121] H. C. Koh and G. Tan, “Data mining applications in healthcare,” *Journal of healthcare information management,* vol. 19, no. 2, pp. 64-72, 2005.

[122] M. K. Obenshain, “Application of data mining techniques to healthcare data,” *Infection control and hospital epidemiology,* vol. 25, no. 8, pp. 690-5, 2004.

[123] D. Tomar, “A survey on Data Mining approaches for Healthcare,” *International Journal of Bio - Science and Bio - Technology,* vol. 5, pp. 241-266, 2013.

[124] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. P. Reinartz, C. Shearer and R. Wirth, “CRISP-DM 1.0: Step-by-step data mining guide,” 2000.

[125] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. J. Ramírez-Quintana and P. Flach, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering,* vol. 33, no. 8, pp. 3048-3061, 2021.

[126] A. Barredo Arrieta, N. Diaz Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado González, S. Garcia, S. Gil-Lopez, D. Molina, V. R. Benjamins, R. Chatila and F. Herrera, "xplainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion,* vol. 58, 2019.

[127] J. Zhu, A. Liapis, S. Risi, R. Bidarra and G. M. Youngblood, "Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation," *2018 IEEE Conference on Computational Intelligence and Games (CIG),* pp. 1-8, 2018.

[128] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. J. Bonten, D. L. Dahly, J. A. Damen, T. P. A. Debray, V. M. T. de Jong and M. a. e. a. De Vos, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ,* vol. 369, 2020.

[129] Alan Turing Institute, "Data science and AI in the age of COVID-19," 2021.

[130] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng and E. Gkrania-Klotsas, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence,* vol. 3, p. 199–217, 2021.

[131] V. Beiranvand, W. Hare and Y. Lucet, "Best Practices for Comparing Optimization Algorithms," *Optimization and Engineering,* vol. 18, 12 2017.

[132] L. Carpinetti and A. de Melo, "What to benchmark? A systematic approach and cases," *Benchmarking: An International Journal,* vol. 9, pp. 244-255, 2002.

[133] C. Molnar, A Guide for Making Black Box Models Explainable, 2 ed., 2022.

[134] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access,* vol. 6, pp. 52138-52160, 2018.

[135] A. Fisher, C. Rudin and F. Dominici, "ll Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," *arXiv,* 2018.

[136] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems,* p. 4768–4777, 2017.

[137] H. Chen, S. Lundberg and S.-I. Lee, "Explaining Models by Propagating Shapley Values of Local Components," *arXiv,* 2019.

[138] C. Olah, A. Mordvintsev and L. Schubert, "Feature visualization," *Distill,* vol. 2, no. 11, 2017.

[139] D. Bau, B. Zhou, A. Khosla, A. Oliva and A. Torralba, “Network Dissection: Quantifying Interpretability of Deep Visual Representations,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3319-3327, 2017.

[140] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *2017 IEEE International Conference on Computer Vision (ICCV),* pp. 618-626, 2017.

[141] Z. Yang, S. Zhang, W. Shen, X. Xing and Y. Gao, “Artificial Intelligence Related Publication Analysis Based on Citation Counting,” *IEEE Access,* vol. PP, 09 2018.

[142] K. Carlson, “Safe Artificial General Intelligence via Distributed Ledger Technology,” *Big Data and Cognitive Computing,* vol. 3, p. 40, 07 2019.

[143] D. Bacciu, P. Lisboa, J. Martín-Guerrero, R. Stoean and A. Vellido, “Bioinformatics and medicine in the era of deep learning,” *ArXiv,* vol. abs/1802.09791, 02 2018.

[144] N. Dey, S. Borra, A. S. Ashour and F. Shi, “Chapter 11 - Application of Machine Learning Algorithms for Classification and Security of Diagnostic Images,” in *Machine Learning in Bio-Signal Analysis and Diagnostic Imaging*, Academic Press, 2019, pp. 273 - 292.

[145] Y. LeCun, Y. Bengio and G. Hinton, “Deep Learning,” pp. 436-44, 05 2015.

[146] G. E. Hinton, S. Osindero and Y.-W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Computation,* vol. 18, p. 1527–1554, 07 2006.

[147] L. Deng and D. Yu, “Deep Learning: Methods and Applications,” *Foundations and Trends in Signal Processing,* vol. 7, p. 197–387, 06 2014.

[148] T. Mikolov, A. Deoras, D. Povey, L. Burget and J. Cernocky, “Strategies for training large scale neural network language models,” *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings,* 12 2011.

[149] A. Krizhevsky, I. Sutskever and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Neural Information Processing Systems,* vol. 25, 01 2012.

[150] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, “Natural Language Processing (Almost) from Scratch,” *The Journal of Machine Learning Research,* vol. 12, p. 2493–2537, 01 2011.

[151] A. Kamilaris and F. Prenafeta Boldú, “A review of the use of convolutional neural networks in agriculture,” *The Journal of Agricultural Science,* vol. 156, pp. 1-11, 06 2018.

[152] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological Cybernetics,* vol. 36, p. 193–202, 1980.

[153] Y. L. Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel and D. Henderson, “Handwritten Digit Recognition with a Back-Propagation Network,” in *Advances in Neural Information Processing Systems 2*, Morgan Kaufmann Publishers Inc., 1990, p. 396–404.

[154] P.-M. Huang and C.-H. Lee, “Estimation of Tool Wear and Surface Roughness Development Using Deep Learning and Sensors Fusion,” *Sensors,* vol. 21, 2021.

[155] A. a. P. B. F. Kamilaris, “A review of the use of convolutional neural networks in agriculture,” *The Journal of Agricultural Science,* vol. 156, pp. 1-11, 06 2018.

[156] R. Yamashita, M. Nishio, R. Do and K. Togashi, “Convolutional neural networks: an overview and application in radiology,” *Insights into imaging,* vol. 9, pp. 611-629, 08 2018 .

[157] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang and Y. Wang, “Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction,” *Sensors,* vol. 17, p. 818, 04 2017.

[158] A. Vo, H. Tran and T. Le, “"Advertisement image classification using convolutional neural network,” in *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, 2017.

[159] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, “Convolutional Neural Networks for Speech Recognition,” *IEEE/ACM Transactions on Audio, Speech and Language Processing,* vol. 22, 10 2014.

[160] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, “Large-scale Video Classification with Convolutional Neural Networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[161] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing,* 08 2014.

[162] A. Khan, A. Sohail, U. Zahoora and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artificial Intelligence Review,* vol. 53, p. 5455–5516, 04 2020.

[163] J. Ou and Y. Li, “Vector-kernel convolutional neural network,” *Neurocomputing,* vol. 300, pp. 253 - 258, 2019.

[164] S. SP, W. L, G. S, G. H, P. P and G. B., “3D Deep Learning on Medical Images: A Review,” *Sensors (Basel),* vol. 20, 09 2020.

[165] E. Kalfaoglu, S. Kalkan and A. Alatan, “Late Temporal Modeling in 3D CNN Architectures with BERT for Action Recognition,” 08 2020.

[166] S. K. Inman, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj and D. J., “1D convolutional neural networks and applications: A survey,” *Mechanical Systems and Signal Processing,* vol. 151, p. 107398, 2021.

[167] A. K. Qureshi, A. Sohail, U. Zahoora and A. Saeed, “A survey of the recent architectures of deep convolutional neural networks,” *Artificial Intelligence Review,* vol. 53, p. 5455–5516, 04 2020.

[168] M. Arsov, E. Zdravevski, P. Lameski, R. Corizzo, N. Koteli, S. Gramatikov, K. Mitreski and V. Trajkovik, “Multi-Horizon Air Pollution Forecasting with Deep Neural Networks,” *Sensors,* vol. 21, 2021.

[169] R. Osman, S. Saleh and Y. Saleh, “A Novel Interference Avoidance Based on a Distributed Deep Learning Model for 5G-Enabled IoT,” *Sensors,* vol. 21, pp. 1424-8220, 2021.

[170] H. Kim, H. Jeong, H. Lee and S. W. Kim, “Online and Offline Diagnosis of Motor Power Cables Based on 1D CNN and Periodic Burst Signal Injection,” *Sensors,* vol. 21, no. 17, 2021.

[171] S. Luo, X. Ma, J. Xu, M. Li and L. Cao, “Deep Learning Based Monitoring of Spatter Behavior by the Acoustic Signal in Selective Laser Melting,” *Sensors,* vol. 21, 2021.

[172] J. Kalajdjieski, E. Zdravevski, R. Corizzo, P. Lameski, S. Kalajdziski, I. M. Pires, N. M. Garcia and V. Trajkovik, “Air Pollution Prediction with Multi-Modal Data and Deep Neural Networks,” *Remote Sensing,* 2020.

[173] S. Albelwi and A. Mahmood, “Framework for Designing the Architectures of Deep Convolutional Neural Networks,” *Entropy,* vol. 19, 06 2017.

[174] J. Snoek, H. Larochelle and R. Adams, “Practical Bayesian Optimization of Machine Learning Algorithms,” *Advances in Neural Information Processing Systems,* vol. 4, 06 2012.

[175] S. R. Young, D. C. Rose, T. P. Karnowski, S.-H. Lim and R. M. Patton, “Optimizing Deep Learning Hyper-Parameters through an Evolutionary Algorithm,” in *MLHPC '15: Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments*, 2015.

[176] B. Z. L and Q. V., “Neural Architecture Search with Reinforcement Learning,” *arXiv,* 2017.

[177] G. Kyriakides and K. G. Margaritis, “An Introduction to Neural Architecture Search for Convolutional Networks,” *arXiv,* 2020.

[178] V. Nekrasov, H. Chen, C. Shen and I. Reid, “Fast Neural Architecture Search of Compact Semantic Segmentation Models via Auxiliary Cells,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 9118-9127, 06 2019.

[179] D. R. S. Le, C. Liang and Q. V., “The Evolved Transformer,” *ArXiv,* 2019.

[180] Pérez-Rúa and J.-M. e. al., “Efficient Progressive Neural Architecture Search,” *British Machine Vision Conference,* 2018.

[181] Google, “Keras: the Python deep learning API,” [Online]. Available: https://keras.io/.

[182] DATA Lab at Texas A&M University, “AutoKeras,” [Online]. Available: https://autokeras.com/. [Accessed (last) 30/12/2020].

[183] H. Jin, Q. Song and X. Hu, “Auto-Keras: An Efficient Neural Architecture Search System,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, NY, USA, 2019.

[184] J. Davison, “DEvol - Deep Neural Network Evolution,” [Online]. Available: https://github.com/joeddav/devol. [Accessed (last) 30/12/020].

[185] Allegro AI, “ClearML - Auto-Magical Suite of tools to streamline your ML workflow Experiment Manager, ML-Ops and Data-Management,” [Online]. Available: https://github.com/allegroai. [Accessed 20 12 (last) 2020].

[186] M. a. O. A. a. B. E. a. S. A. a. B. M. a. L. C. de Onis, “Worldwide implementation of the WHO Child Growth Standards,” *Public Health Nutrition,* vol. 15, p. 1603–1610, 2012.

[187] C. a. A. L. a. F. C. a. H. P. a. M. R. a. R. L. a. S. H. Victora, “Maternal and child undernutrition. Consequences for adult health and human capital,” vol. 371, pp. 340-57, 02 2008.

[188] E. a. N. S. Ziegler, “The WHO growth standards: Strengths and limitations,” *Current opinion in clinical nutrition and metabolic care,* vol. 15, pp. 298-302, 02 2012.

[189] J. M. Tanner, “Growth as a Mirror of the Condition of Society: Secular Trends and Class Distinctions,” *Pediatrics International,* vol. 29, pp. 96-103, 1987.

[190] J. M. Tanner, “Normal growth and techniques of growth assessment,” *Clinics in Endocrinology and Metabolism,* vol. 15, pp. 411 - 451, 1986.

[191] S. M. Garn, “Physical growth and development,” *American Journal of Physical Anthropology,* vol. 10, pp. 169-192, 1952.

[192] O. a. T. L. a. H. C. Bilukha, “Impact of New WHO Growth Standards on the Prevalence of Acute Malnutrition and Operations of Feeding Programs-Darfur, Sudan,” *JAMA The Journal of the American Medical Association,* vol. 302, pp. 484-485, 2009.

[193] A. a. K. M. Seal, “Operational implications of using 2006 World Health Organization growth standards in nutrition programmes: secondary data analysis,” *BMJ,* vol. 334, p. 733, 2007.

[194] World Health Organization, “ WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development,” 2006.

[195] Y. a. M. L. a. C. B. a. C. T. Wang, “Limitations of the Current World Health Organization Growth References for Children and Adolescents,” *Food and nutrition bulletin,* vol. 27, pp. S175-88, 01 2007.

[196] E. a. d. O. M. a. G. C. a. V. d. B. J. a. F. E. A. a. G.-S. L. a. V. B. S. a. P. H. a. M. L. a. M. R. a. O. A. W. a. M. J. C. Borghi, “Construction of the World Health Organization child growth standards: selection of methods for attained growth curves,” *Statistics in Medicine,* vol. 25, pp. 247-265, 2006.

[197] World Health Organization, “The WHO Multicentre Growth Reference Study (MGRS),” 2003.

[198] N. a. H. N. Cameron, “Should the UK use WHO growth charts?,” *Paediatrics and Child Health,* vol. 20, pp. 151-156, 04 2010.

[199] M. W. P. a. S. Szefler, “Personalized medicine in children with asthma,” *Paediatric Respiratory Reviews,* vol. 16, pp. 101-107, 2015.

[200] K. a. B. C. G. Cornetta, “Balancing personalized medicine and personalized care,” *Academic medicine : journal of the Association of American Medical Colleges,* pp. 309-313, 03 2013.

[201] I. S. a. G. G. S. Chan, “Personalized Medicine: Progress and Promise,” *Annual Review of Genomics and Human Genetics,* vol. 12, pp. 217-244, 2011.

[202] H. Gray, “Prediction of Adult Stature,” *Child Development,* vol. 19, pp. 167-175, 1948.

[203] J. M. G. H. W. R. H. Tanner, “Standards for children's height at ages 2-9 years allowing for heights of parents,” *Archives of disease in childhood,* pp. 755-762, 1970.

[204] C. a. C. T. Wright, “The strengths and limitations of parental heights as a predictor of attained height,” *Archives of disease in childhood,* vol. 81, pp. 257-60, 10 1999.

[205] T. a. W. C. {Cole, “A chart to predict adult height from a child's current height,” *Annals of human biology,* pp. 662-8, 07 2011.

[206] J. M. a. W. R. H. a. M. W. A. a. C. B. S. Tanner, “Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4 to 16 with allowance for midparent height,” *Archives of Disease in Childhood,* vol. 50, pp. 14-26, 1975.

[207] J. a. L. K. a. C. N. a. C. B. a. P. J. Tanner, “Prediction of adult height from height and bone age in childhood: A new system of equations (TW Mark II) based on a sample including very tall and very short children,” *Archives of disease in childhood,* vol. 58, pp. 767-76, 11 1983.

[208] M. a. M. S. Hossin, “A Review on Evaluation Metrics for Data Classification Evaluations,” *International Journal of Data Mining & Knowledge Management Process,* vol. 5, pp. 01-11, 03 2015.

[209] P. Flach, “The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics,” *Proceedings of ICML03,* pp. 194-201, 01 2003.

[210] M. a. J. N. a. S. S. Sokolova, “Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation,” *AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science,* vol. 4304, pp. 1015-1021, 01 2006.

[211] E. a. E. H. a. E. H. I. Namli, “Artificial Intelligence-Based Prediction Models for Energy Performance of Residential Buildings,” in *Recycling and Reuse Approaches for Better Sustainability*, Springer International Publishing, 2019, pp. 141-149.

[212] F. Galton, “Anthropometric percentiles,” *Nature,* vol. 31, pp. 223-5, 1885.

[213] A. L. W. a. G. E. D. Wachsmuth, “Galton's Bend: A Previously Undiscovered Nonlinearity in Galton's Family Stature Regression Data,” *The American Statistician,* vol. 57, pp. 190-192, 2003.

[214] E. a. V. E. Frees, “Understanding Relationships Using Copulas,” *North American Actuarial Journal,* vol. 2, pp. 1-25, 01 1998.

[215] M. A. Z. a. J. N. S. A. N. Chadha, “Optimal feature extraction and selection techniques for speech processing: A review,” *2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, 2016,* pp. 1669-1673.

[216] J. a. A. S. a. L. H. Tang, “Feature selection for classification: A review,” *Data Classification: Algorithms and Applications,* pp. 37-64, 01 2014.

[217] I. I. P. L. Yvan Saeys, “A review of feature selection techniques in bioinformatics,” *Bioinformatics,* vol. 23, p. 2507–2517, October 2007.

[218] X. a. K. V. a. R. Q. J. a. G. J. a. Y. Q. a. M. H. a. M. G. J. a. N. A. a. L. B. a. Y. P. S. a. Z. Z.-H. a. S. M. a. H. D. J. a. S. Wu, “Top 10 Algorithms in Data Mining,” *Knowledge and Information Systems,* vol. 14, pp. 1-37, 12 2007.

[219] P. a. L. C. W. a. M. J. H. Orzechowski, “Where Are We Now?: A Large Benchmark Study of Recent Symbolic Regression Methods,” *Proceedings of the Genetic and Evolutionary Computation Conference,* pp. 1183-1190, 2018.

[220] R. a. L. G. Brereton, “Support Vector Machines for classification and regression,” *The Analyst,* vol. 135, pp. 230-67, 02 2010.

[221] H. a. C. C. a. K. L. a. S. A. a. V. V. Drucker, “Support Vector Regression Machines,” *Advances in Neural Information Processing Systems,* vol. 9, 11 2003.

[222] G. a. M. Q. a. F. T. a. W. T. a. C. W. a. M. W. a. Y. Q. a. L. T.-Y. Ke, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” *Proceedings of the 31st International Conference on Neural Information Processing Systems,* pp. 3149-3157, 2017.

[223] T. a. G. C. Chen, “XGBoost: A Scalable Tree Boosting System,” *The 22nd ACM SIGKDD International Conference,* pp. 785-794, 08 2016.

[224] E. A. Daoud, “Comparison between XGBoost , LightGBM and CatBoost Using a Home Credit Dataset,” *International Journal of Computer and Information Engineering,* vol. 13, 2019.

[225] R. Hafizur, T. Tasmia Jannat, A. Syed Mithun and P. Sanjoy Kumar, “A grey approach to predicting healthcare performance,” *Measurement,* vol. 134, pp. 307-325, 2019.

[226] P. Thungjaroenkul, G. Cummings and A. Embleton, “The Impact of Nurse Staffing on Hospital Costs and Patient Length of Stay,” *Nursing Economic$,* vol. 25, pp. 255-65, 2007.

[227] F. Ma, L. Yu, L. Ye, D. D. Yao and W. Zhuang, “Length-of-Stay Prediction for Pediatric Patients With Respiratory Diseases Using Decision Tree Methods,” *IEEE Journal of Biomedical and Health Informatics,* vol. 24, no. 9, pp. 2651-2662, 2020.

[228] M. Berry, P. Shah and R. e. a. Brouillette, “Predictors of mortality and length of stay for neonates admitted to children's hospital neonatal intensive care units.,” *Journal of Perinatology,* vol. 28, p. 297–302, 2008.

[229] B. Zernikow, K. Holtmannspötter and E. Michel, “Predicting length-of-stay in preterm neonates,” *European Journal of Pediatrics,* vol. 158, p. 59–62, January 1999.

[230] T. Zhu, L. Luo, X. Zhang and W. Shen, “Modeling the Length of Stay of Respiratory Patients in Emergency Department Using Coxian Phase-Type Distributions With Covariates,” *IEEE Journal of Biomedical and Health Informatics,* vol. 22, no. 3, pp. 955-965, 2018.

[231] J. M. P. Gutiérrez, M.-Á. Sicilia, S. Sanchez-Alonso and E. García-Barriocanal, “Predicting Length of Stay Across Hospital Departments,” *IEEE Access,* vol. 9, pp. 44671-44680, 2021.

[232] S. Shamsir, O. Hassan and S. K. Islam, “Smart Infant-Monitoring System with Machine Learning Model to Detect Physiological Activities and Ambient Conditions,” *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC),* pp. 1-6, 2020.

[233] M. S. Kamal, A. Northcote, L. Chowdhury, N. Dey, R. G. Crespo and E. Herrera-Viedma, “Alzheimer’s Patient Analysis Using Image and Gene Expression Data and Explainable-AI to Present Associated Genes,” *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-7, 2021.

[234] Z. Liu, H. Wang, Y. Gao and S. Shi, “Automatic Attention Learning Using Neural Architecture Search for Detection of Cardiac Abnormality in 12-Lead ECG,” *IEEE Transactions on Instrumentation and Measurement,* vol. 70, pp. 1-12, 2021.

[235] L. M. Mahieu, N. Buitenweg, P. Beutels and J. J. De Dooy, “Additional hospital stay and charges due to hospital-acquired infections in a neonatal intensive care unit,” *The Journal of hospital infection,* vol. 47, no. 3, p. 223–229, 2001.

[236] J. Ramasethu, “Prevention and treatment of neonatal nosocomial infections,” *Maternal health, neonatology and perinatology,* vol. 3, February 2017.

[237] J. Shepard, J. Frederick, F. Wong, S. Madison, L. Tompkins and E. Hadhazy, “Could the prevention of health care-associated infections increase hospital cost? The financial impact of health care-associated infections from a hospital management perspective,” *American journal of infection control,* vol. 48, no. 3, p. 255–260.

[238] World Health Organization, “Report on the burden of endemic health care-associated infection worldwide,” Geneve, 2011.

[239] Robert Koch-Institut, “Nationales Referenzzentrum für Surveillance von nosokomialen Infektionen,” [Online]. Available: https://www.nrz-hygiene.de/en/surveillance/hospital-infection-surveillance-system/neo-kiss/. [Accessed 15 02 2021].

[240] P. Powell, C. Powell, S. Hollis and M. Robinson, “When will my baby go home?,” *Archives of disease in childhood,* vol. 67, pp. 1214-6, 11 1992.

[241] A. L. Brodin, A. Bedu, P. Lombrail, L. Desplanques, B. Diakite, E. Bingen and Y. Aujard, “Prolongation of hospital stay and extra costs due to hospital-acquired infection in a neonatal unit,” *Journal of Hospital Infection,* vol. 35, no. 1, pp. 37 - 45, 1997.

[242] C. Phibbs and S. Schmitt, “Estimates of the Cost and Length of Stay Changes that can be Attributed to One-Week Increases in Gestational Age for Premature Infants,” *Early human development,* vol. 82, pp. 85-95, 02 2006.

[243] S. Kieran, R. Zwiggelaar, P. Jones and N. Mac Parthaláin, “A systematic review of the prediction of hospital length of stay: Towards a unified framework,” *PLOS Digital,* vol. 1, no. 4, 2022.

[244] H. Xie, T. Chaussalet and P. Millard, “A Model-Based Approach to the Analysis of Patterns of Length of Stay in Institutional Long-Term Care,” *IEEE Transactions on Information Technology in Biomedicine,* vol. 10, no. 3, pp. 512-518, 2006.

[245] H. Li, H. Gong, S. Gu and K. Chen, “Grey correlation analysis of Average Length of Stay in general hospitals,” *2020 Chinese Control And Decision Conference (CCDC),* pp. 4102-4107, 2020.

[246] K. M, Y. N, S. C and G. N., “Cost Analysis on Intensive Care Unit Costs Based on the Length of Stay,” *Turkish journal of anaesthesiology and reanimation,* vol. 47, no. 2, pp. 142-145, 2019.

[247] S. Grampurohit and S. Sunkad, “Hospital Length of Stay Prediction using Regression Models,” *2020 IEEE International Conference for Innovation in Technology (INOCON),* pp. 1-5, 2020.

[248] A. Kumar and H. Anjomshoa, “A Two-Stage Model to Predict Surgical Patients’ Lengths of Stay From an Electronic Patient Database2019,” *IEEE Journal of Biomedical and Health Informatics,* vol. 23, no. 2, pp. 848-856, 2019.

[249] Y. Weng, T. Zhou, Y. Li and X. Qiu, “NAS-Unet: Neural Architecture Search for Medical Image Segmentation,” *IEEE Access,* vol. 7, pp. 44247-44257, 2019.

[250] A. Kwasigroch, M. Grochowski and A. Mikołajczyk, “Neural Architecture Search for Skin Lesion Classification,” *IEEE Access,* vol. 8, 2020.

[251] C. D, W. MJ and J. G, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” *PeerJ. Computer science,* vol. 7, 2021.

[252] A. Vimont, H. Leleu and I. Durand-Zaleski, “Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database,” *The European Journal of Health Economics,* vol. 23, p. 211–223, 2022.

[253] d. G. P. Direção-Geral da Saúde, “Programa de Prevenção e Controlo de Infeções e de Resistência aos Antimicrobianos (PPCIRA),” [Online]. Available: https://www.dgs.pt/programa-de-prevencao-e-controlo-de-infecoes-e-de-resistencia-aos-antimicrobianos/apresentacao.aspx.

[254] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu.

[255] F. Alzamzami, M. Hoda and A. E. Saddik, “Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation,” *IEEE Access,* vol. 8, 2020.

[256] S. Caprio, N. Santoro and R. Weiss, “Childhood obesity and the associated rise in cardiometabolic complications,” *{Nature Metabolism,* vol. 2, pp. 223-232, 03 2020.

[257] World Health Organization, “Taking action on childhood obesity report,” World Health Organization, 2018.

[258] M. Blüher, “Obesity: global epidemiology and pathogenesis,” *Nature Reviews Endocrinology,* vol. 15, no. 5, p. 288–298, 2019.

[259] World Health Organization, “WHO - Obesity,” April 2021. [Online]. Available: https://www.who.int/health-topics/obesity.

[260] National Heart, Lung, and Blood Institute, “Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: The Evidence Report,” 1998.

[261] T. a. B. L. a. U. R. Lobstein, “Obesity in children and young people: A crisis in public health,” *Obesity reviews: an official journal of the International Association for the Study of Obesity ,* pp. 5-104, 01 2004.

[262] R. JJ, M. E, M. ZC, H. B, A. D, S. L and K. CJ, “Health consequences of obesity,” *Archives of disease in childhood,* vol. 88, no. 9, pp. 748-752, 2003.

[263] D. S. Freedman, L. Kettel Khan, M. K. Serdula, S. R. Srinivasan and G. S. Berenson, “BMI rebound, childhood height and obesity among adults: the Bogalusa Heart Study,” *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity,* vol. 25, no. 4, p. 543–549, 2001.

[264] Z. Saboia, G. Meneses, A. Martins, E. Daher and G. Silva Junior, “Association between syndecan-1 and renal function in adolescents with excess weight: Evidence of subclinical kidney disease and endothelial dysfunction,” *Brazilian Journal of Medical and Biological Research,* vol. 51, 01 2018.

[265] A. Costa, L. Correia-Costa, A. Afonso, F. Schaefer, A. Guerra, C. Moura, C. Mota, H. Barros, J. Areias and A. Azevedo, “Determinants of carotid-femoral pulse wave velocity in prepubertal children,” *International Journal of Cardiology,* vol. 218, 05 2016.

[266] S. Weihrauch-Blüher, P. Schwarz and J. H. Klusmann, “Childhood obesity: increased risk for cardiometabolic disease and cancer in adulthood,” *Metabolism: clinical and experimental,* vol. 92, pp. 147-152, 2019.

[267] P. Poirier and R. H. Eckel, “Obesity and cardiovascular disease,” *Current Atherosclerosis Reports,* vol. 4, no. 6, p. 448–453, 2002.

[268] Avon Longitudinal Study of Parents and Children Study Team, “Early life risk factors for obesity in childhood: cohort study,” *BMJ (Clinical research ed.),* vol. 330, no. 7504, 2005.

[269] R. M. Krauss, M. Winston, B. J. Fletcher and S. M. Grundy, “Obesity : impact on cardiovascular disease. Circulation,” *Circulation,* vol. 98, no. 14, p. 1472–1476, 1998.

[270] C. Paech, J. Moser, I. Dähnert, F. Wagner, R. A. Gebauer, T. Kirsten, M. Vogel, W. Kiess, A. Körner, B. Wolfarth and J. Wüstenfeld, “Different habitus but similar electrocardiogram: Cardiac repolarization parameters in children - Comparison of elite athletes to obese children,” *Annals of pediatric cardiology,* vol. 12, no. 3, p. 201–205, 2019.

[271] S. Genovesi, P. Salvi, E. Nava and e. a. Elena Tassistro, “Blood Pressure and Body Weight Have Different Effects on Pulse Wave Velocity and Cardiac Mass in Children,” *Journal of Clinical Medicine,* vol. 9, no. 9, p. 2954, 2020.

[272] G.-Z. Sun, Y. Li, X.-H. Zhou, X.-F. Guo, X.-G. Zhang, L.-Q. Zheng, Y. Li, Y.-D. Jiao and Y.-X. Sun, "Association between obesity and ECG variables in children and adolescents: A cross-sectional study," *Experimental and therapeutic medicine,* vol. 6, pp. 1455-1462, 12 2013.

[273] M. Alpert, B. Terry, C. Hamm, T.-H. Fan, M. Cohen, C. Massey and J. Painter, "Effect of Weight Loss on the ECG of Normotensive Morbidly Obese Patients," *Chest,* vol. 119, pp. 507-10, 02 2001.

[274] S. M. Grundy, "Metabolic Syndrome, and Cardiovascular Disease," *The Journal of Clinical Endocrinology & Metabolism,* vol. 89, no. 6, p. 2595–2600, 2004.

[275] M. Faienza, M. Chiarito and E. e. a. Molina-Molina, "Childhood obesity, cardiovascular and liver health: a growing epidemic with age," *World Journal of Pediatrics,* vol. 16, no. 5, p. 438–445, 2020.

[276] G. Twig, G. Yaniv, H. Levine, A. Leiba, N. Goldberger, E. Derazne, D. Ben-Ami Shor, D. Tzur, A. Afek, A. Shamiss, Z. Haklai and J. D. Kark, " Body-Mass Index in 2.3 Million Adolescents and Cardiovascular Death in Adulthood," *The New England journal of medicine,* vol. 374, no. 25, pp. 2430-2440, 2016.

[277] Departmento de Epidemiologia Clinica, Medicina preditiva e Saúde pública da Universidade do Porto, "Geração 21," [Online]. Available: https://www.geracao21.com/. [Accessed 2021].

[278] P. Madonaa, R. Ilias, B. Muhammad and M. Zain, "PQRST wave detection on ECG signals," *Gaceta Sanitaria,* vol. 35, pp. S364-S369, 2021.

[279] L. Richardson, "Beautiful Soup Documentation," [Online]. Available: https://beautiful-soup-4.readthedocs.io/en/latest/#.

[280] M. de Onis, A. W. Onyango, E. Borghi, A. Siyam, C. Nishida and J. Siekmann, "Development of a WHO growth reference for school-aged children and adolescents," *Bulletin of the World Health Organization,* vol. 85, no. 9, p. 660–667, 2007.

[281] M. G. Siqueira and P. S. R. Diniz, "2 - Digital Filters," in *The Electrical Engineering Handbook*, Academic Press, 2005, pp. 839-860.

[282] D. Gunning, "DARPA's explainable artificial intelligence (XAI) program," *IUI '19: Proceedings of the 24th International Conference on Intelligent User Interfaces,* 2019.

[283] N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion,* vol. 58, pp. 82-115, 2020.

[284] A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV),* pp. 839-847, 2018.

[285] Y. a. B. Y. a. H. G. LeCun, "Deep Learning," pp. 436-44, 05 2015.

[286] L. a. Y. D. Deng, “Deep Learning: Methods and Applications,” *Foundations and Trends in Signal Processing,* vol. 7, p. 197–387, 06 2014.

[287] T. a. D. A. a. P. D. a. B. L. a. C. J. Mikolov, “Strategies for training large scale neural network language models,” *2011 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2011, Proceedings,* 12 2011.

[288] A. a. S. I. a. H. G. Krizhevsky, “ImageNet Classification with Deep Convolutional Neural Networks,” *Neural Information Processing Systems,* vol. 25, 01 2012.

[289] R. a. W. J. a. B. L. a. K. M. a. K. K. a. K. P. Collobert, “Natural Language Processing (Almost) from Scratch,” *The Journal of Machine Learning Research,* vol. 12, p. 2493–2537, 01 2011.

[290] Y. L. a. B. B. a. D. J. S. a. H. R. E. a. H. W. a. J. L. D. a. H. D. Cun, “Handwritten Digit Recognition with a Back-Propagation Network,” in *Advances in Neural Information Processing Systems 2*, Morgan Kaufmann Publishers Inc., 1990, p. 396–404.

[291] C. M. Curry, J. D. Ross, A. J. Contina and E. S. Bridge, “Varying dataset resolution alters predictive accuracy of spatially explicit ensemble models for avian species distribution,” *Ecology and Evolution,* vol. 8, no. 24, pp. 12867-12878, 2018.

[292] M. H. Hashem, M. M. Islam, M. I. I. Zarif and M. Hashem, “Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches,” *Internet of Things,* vol. 7, p. 100059, 2019.

[293] A. Rosenberg and J. Hirschberg, “V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure.,” *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* pp. 410-420, 2007.

[294] J.-O. Palacio-Niño and F. Berzal, “Evaluation Metrics for Unsupervised Learning Algorithms,” *arXiv,* 2019.

Appendix