UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR
DE INGENIEROS DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA BIOMÉDICA

TRABAJO FIN DE GRADO

DESIGN AND DEVELOPMENT OF AN AUTOMATED SYSTEM FOR EARLY DETECTION OF ELDERLY FRAILTY

JULIA DE ENCISO GARCÍA
JUNIO 2024

TRABAJO DE FIN DE GRADO

Título:	Diseño y Desarrollo de un Sitema Automático para la Detección Precoz de Fragilidad en Ancianos
Título (inglés):	Design and Development of an Automated System for Early Detection of Elderly Frailty
Autor:	Julia de Enciso García
Tutor:	Óscar Araque Iborra
Departamento:	Departamento de Ingeniería de Sistemas Telemáticos

MIEMBROS DEL TRIBUNAL CALIFICADOR

Presidente:	—
Vocal:	—
Secretario:	—
Suplente:	—

FECHA DE LECTURA:

CALIFICACIÓN:

UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE
INGENIEROS DE TELECOMUNICACIÓN

Departamento de Ingeniería de Sistemas Telemáticos
Grupo de Sistemas Inteligentes

TRABAJO FIN DE GRADO

DESIGN AND DEVELOPMENT OF AN AUTOMATED SYSTEM FOR EARLY DETECTION OF ELDERLY FRAILTY

Julia de Enciso García

Junio 2024

Resumen

El envejecimiento de la población y el consiguiente aumento de la esperanza de vida suponen un impacto significativo en los sistemas sanitarios, especialmente en relación con la fragilidad mental en personas mayores. Este incremento en la longevidad conlleva mayores costos para el sector sanitario, así como una creciente demanda de atención en salud mental relacionada con el envejecimiento, lo que ejerce una presión sobre los recursos sanitarios disponibles.

El objetivo principal de este trabajo es desarrollar un sistema de aprendizaje automático que asista a los clínicos en la detección y predicción de la fragilidad mental en sus etapas más tempranas. Inicialmente, para lograrlo, se llevó a cabo una extracción de las características con más relevancia, permitiendo reducir la cantidad de evaluaciones a las que un paciente debe someterse. Esto no solo alivia la carga de trabajo de los clínicos al eliminar la recopilación de datos no influyentes, sino que también ofrece la posibilidad de detectar la fragilidad sin necesidad de tantas pruebas.

Por otro lado, los modelos desarrollados muestran buenas capacidades en la predicción de fragilidad mental, ya que permiten la extracción de relaciones entre características relevantes y la exploración y desarrollo de modelos mediante diversas alternativas de aprendizaje automático, como el perceptrón multicapa y las redes convolucionales (CNN).

Tras la exploración, se evaluaron los modelos desarrollados para mejorar el rendimiento del sistema, centrándose en el F1-score y apoyando las estrategias de prevención junto con las pruebas de diagnóstico, como el FFP. Además, se prestó especial atención a las dependencias temporales entre los datos recopilados en diferentes momentos para comprender la evolución de la fragilidad mental y ofrecer una atención más personalizada y efectiva.

En cuanto a los resultados obtenidos, se compararon los diferentes algoritmos con los recogidos por el estudio de Leghissa [1], destacando aquellos que han demostrado un mejor rendimiento en la detección y predicción de la fragilidad mental en etapas tempranas. Estas conclusiones proporcionan una base sólida para futuras investigaciones y para la implementación de sistemas de apoyo a la toma de decisiones clínicas en este ámbito.

Palabras clave: Aprendizaje Automático, fragilidad, Redes Neuronales, Redes Convolucionales, envejecimiento, Recursive Feature Extraction, perceptron multicapa

Abstract

The ageing of the population and the resulting increase in life expectancy have a significant impact on countries' healthcare systems, especially with regard to mental frailty of older people. This increase in longevity leads to higher costs of medical care, as well as an increase demand for mental health care problems related to ageing, putting additional pressure on available health resources.

The main objective of this study is to develop a machine learning system that can assist clinicians in the early detection and prediction of mental frailty. Initially, in order to achieve this objective, an extraction of the most relevant features was performed, which allowed for a reduction in the number of assessments a patient has to undergo. This not only alleviates the burden on clinicians by eliminating irrelevant data collection, but also enables the detection of frailty without the necessity of so many tests.

Conversely, the models developed demonstrate satisfactory capabilities in the prediction of mental frailty, as they permit the extraction of relationships between relevant features and the development of models using various machine learning alternatives, such as Multilayer Perceptron (MLP) and convolutional networks (CNN).

After the exploration, the models developed were evaluated according to their F1-score, with the aim of identifying and improving system performance and supporting prevention strategies, in conjuction with other diagnostic tests, such as FFP. Furthermore, particular emphasis was placed on the incorporation of temporal dependencies between data collected at different times. This approach allows a better understanding of the evolution of mental frailty over time, contributing to a more personalised and effective treatment.

In light of the outcomes observed, a number of algorithms were evaluated in comparison to those derived from the study conducted by Leghissa [1]. This analysis has identified those algorithms and models that have demonstrated superior performance in the detection and prediction of mental frailty in its early stages. These findings provide a robust foundation for future research and the implementation of clinical decision support systems in this area.

Keywords: Machine Learning, frailty, ageing, Neural Netork, Convolutional Network, Recursive Feature Extraction, Multilayer Perceptron

Agradecimientos

En primer lugar, dar las gracias a mi tutor Óscar, por su disponibilidad, paciencia y, especialmente, por su incesante confianza. También agradecer a Matteo, cuya valiosa ayuda y colaboración han permitido que este trabajo saliese a la luz.

Pero, principalmente, agradecer a mis padres, por inculcarme que *no era suerte, sino trabajo, esfuerzo y ganas*, y que lo primero, siempre será la familia, a pesar de que *lo primero es lo primero*.

Contents

List of Figures

CHAPTER 1

Introduction

1.1 Context

Global population ages and every country in the world is experiencing an striking growth in the proportion of older people and its pace is becoming faster. This demographic shift poses a major challenge for their health and social systems since, 9,82% of the world population is over 65 years old and, specifically, about 20,20% of Spanish populace belongs to this age group [3]. According to WHO [4], 1 in 6 people in the world will be aged 60 years or over by 2030.

Despite the fact that people live longer, their extended life expectancy leads to a deterioration of both, physical and mental capacities, frequently related to frailty. The identification and management of frailty among other adults have become increasingly crucial in geriatric medicine, since early detection and early prediction techniques would assist in the management of age-related conditions, thereby reducing the necessity for costly and invasive treatment.

Frailty progression significantly increases the utilization of health resources. In Catalonia, Spain, the prevalence of frailty leads to a 125% rise in healthcare costs, primarily attributed to hospitalizations. This corresponds to an additional annual healthcare expendi-

ture of approximately 1170€ per frail individual compared to their non-frail counterparts [5].

Frailty, characterized by age-associated declines in physiologic reserve and function across multiple organ systems, poses significant risks for adverse health outcomes, including falls, disability, hospitalization, and mortality [6, 7]. Despite its clinical significance, defining frailty has been a challenge until recent years.

Various conceptualizations of frailty exist in the literature. One widely recognized model, proposed by Fried et al. [8], delineates frailty as a clinical phenotype (FFP) characterized by a cycle of negative energy balance, sarcopenia, and diminished strength and tolerance for exertion [9]. This phenotype identifies frailty as a state of increased vulnerability resulting from aging-associated declines in reserve and function, impairing the ability to cope with every day or acute stressors.

Alternatively, frailty has been quantified through a risk index known as the "frailty index (FI)" [10], which tallies the accumulation of deficits over time, encompassing disability, illnesses, physical and cognitive impairments, psychosocial risk factors, and geriatric syndromes. It has been contended that, in contrast to Fried's frailty phenotype, the FI serves as a more sensitive predictor of adverse health outcomes due to its finely graded risk scale [6].

Operational definitions of frailty, such as Fried's phenotypic criteria or frailty index model, aim to identify individuals at risk of adverse health outcomes by assesing factors like grip strength, slow walking speed, energy, low physical activity, and unintentional weight loss. Considering a pre-frail person when one or two criteria are present, and in the frail stage when three out of five phenotypic criteria are met[11].

Moreover, frailty is also considered to have a multidimensional nature. It is a combination of different dimensions on the level of physical, psychological or social condition of the elder [12]. All of them are in interaction with each other while exerting a mutual influence upon each other. Likewise, elders' frail status depends on external factor such as the environment. Physical frailty primarly encompasses the physiological decline and, conversely, psychological frailty pertains to emotional, cognitive, and psychosocial aspects. While physical frailty underscores the somatic manifestations of decline, psychological frailty illuminates the intricate interplay between mental and emotional well-being, emphasizing the importance of addressing holistic aspects of health and resilience in geriatric care and intervention strategies [13, 14].

Additionally, efforts have been made to develop consensus on frailty, acknowledging it as a clinical syndrome indicative of increased vulnerability, reversible with interventions, and

useful in primary care settings. Despite these advancements, challenges persist in frailty research and practice. Population aging exacerbates the prevalence of frailty, necessitating effective strategies for early detection and intervention.

Research on frailty, particularly of its earliest stages, which are known as pre-frailty, is of crucial importance for improving health outcomes among older adults. Among the various tools and technologies utilised in healthcare, numerous studies have highlighted the potential of Artificial Intelligence (AI) in the identification of early signs of frailty. It is noteworthy that Machine Learning (ML), a branch of AI, has emerged as a powerful instrument in medical research, assisting clinical staff in making informed decisions [15].

Machine Learning is adept at predicting and understanding complex patterns derived from extensive datasets. The utilisation of data collected through advanced technologies, such as sensors [16], enables the application of ML algorithms to provide significant insights aimed at identifying frailty [17]. This superiority is due to the superior performance of ML compared to traditional analytical methods. It excels in uncovering intricate relationships within data and managing large, diverse datasets, which are typical in healthcare. This encompasses the capacity to handle the inherent intra- and inter-patient variability, to learn from new data over time, and to adapt to changes.

Furthermore, Machine Learning is capable of effectively managing incomplete or invalid data, thereby mitigating the impact of such data on the results. This is achieved through the utilisation of sophisticated algorithms that employ feature selection and dimensionality reduction techniques, thereby revealing patterns within the data that would otherwise remain hidden. These capabilities render Machine Learning an invaluable tool in the early detection of frailty, thereby contributing to improved health management and quality of life for older adults.

In light of the aforementioned challenges, there are projects such as MIRATAR [18], which focuses on the development of intelligent systems for the early detection and prevention of mental health problems in older people. This work is embedded in the context of this project, which seeks to address the challenges of population ageing and its impact on health systems, using advanced technologies to improve the quality of life of older adults and reduce the burden on health resources.

Nevertheless, the lack of a universal quantitative definition of frailty complicates comparisons across studies and integration of these advancements in Machine Learning and the developments of standardised assessment tools.

1.2 Project goals

The primary objective of this project is to develop an automated system using Machine Learning (ML) algorithms for the early detection of mental frailty, as this geriatric syndrome is currently emerging as one of the most important scientific issues in gerontology. This is to be done based on the Fried Frailty Phenotype (FFP). The objectives can be listed as follows:

1. Use feature extraction methods with the aim of reducing the number of tests to be performed on patients and incorporate them as features in the algorithms.
2. Explore and develop model alternatives with different strategies to improve system performance.
3. Evaluation the system performance to aid in the prevention strategies alongside other diagnostic tests, such as FFP.

1.3 Structure of this document

In this section we provide a brief overview of the chapters included in this document. The structure is as follows:

Chapter 1 presents a general context and summary of the project, as well as the objectives to be achieved. The aim is to elucidate the rationale driving this research, the challenges it aims to tackle, and the methodologies employed to effectively address them.

Chapter 2 provides information about the tools and technologies used throughput the project, as well as similar studies intended to face the same problem.

Chapter 3 specifies the different models, the motivation for their use and, how they will be developed.

Chapter 4 explains where data was obtained and introduces all results obtained from the models.

Chapter 5 concludes by presenting findings drawn directly form the research, summarizing achieved objectives, and engaging in a discussion regarding potential avenues for future development and improvement of this work.

Chapter A and Chapter B discusses the impact and economic budget of the project.

CHAPTER 2

Background

This chapter will outline the different enabling technologies and relevant research related to the project. It will introduce the programming language employed, Python, along with the primary libraries utilised and the development environment. Lastly, the chapter will delve into Machine Learning technologies. The related work section will examine prior research concerning the prediction and detection of frailty and the use of ML in medicine.

2.1 Enabling technologies

2.1.1 Python

This project has been implemented using Pyhton as the main programming language. Python is a high-level, object-oriented and readable programming language known for its simplicity [19]. It's widely used due to its extensive collection of libraries, specifically designed for data science and machine learning applications. Following, the primary libraries employed will be described.

Initially, **Pandas** [20], an open-source library, a go-to tool for data analysis and manipulation utilised extensively. With its intuitive interface, it offers efficient handling of large

datasets through features like lazy evaluation and memory optimization. Its key functionalities include managing data structures like Series and DataFrame, facilitating data cleaning and preprocessing, and seamless integration with other data analysis libraries.

Additionally, for the graphical representations in this work, **Matplotlib** [21], another open-source Python library, is a powerful tool for creating static, interactive, and publication-quality visualizations. It provides a wide range of plotting functions and customization options, making it ideal for various data visualization tasks. On the other hand, **Seaborn** [22], built on top of Matplotlib, offers a higher-level interface for creating attractive and informative statistical graphics. It simplifies the process of producing complex visualizations by providing easy-to-use functions for common statistical plots.

Together, Matplotlib and Seaborn form a formidable duo for data visualization in Python. While Matplotlib provides a foundation for creating plots with extensive customization, Seaborn enhances this capability with its streamlined interface and specialised functions for statistical visualization.

Furthermore, **NumPy** [23] is the cornerstone of numerical computing in Python, providing efficient tools for working with arrays, matrices, and mathematical functions. It enables developers to perform a wide range of operations, from basic array manipulation to advanced mathematical computations. With its comprehensive functionality, NumPy is essential for tasks such as scientific computing, data analysis, and machine learning.

With NumPy's rich functionality, developers can handle large datasets with ease, execute complex algorithms, and solve intricate mathematical problems. In essence, NumPy empowers users with the computational muscle needed to tackle diverse challenges, driving advancements in research, engineering, and data-driven decision-making.

In addition, **Pickle** [24] is a Python module used for the serialisation and de-serialisation of Python objects. It enables developers to convert complex objects into byte streams for storage or transmission and reconstruct them later. The utilisation of this module throughout the work has facilitated the efficient storage of various elements, simplifying data management and enabling seamless integration across different parts of the project.

Lastly, the development environment adopted was **Jupyter Notebook** [25], an interactive web-based platform. In addition to facilitating code execution, Jupyter Notebook offers an user-friendly interface that facilitates rapid prototyping and experimentation, enabling iteratively develop and document their code and analysis with comprehensive media representation of results and supports markup language. The notebooks are made of cells that can be executed independently. The notebooks contain cells that can be executed sep-

arately, allowing the code to be run incrementally and providing immediate visualization of results.

2.1.2 Machine Learning

During the project's development, machine learning algorithms were applied using the Python libraries Scikit-learn and Tensorflow.

Firstly, **Scikit-learn** [26] is a Python library for machine learning, providing a wide range of tools for data mining and analysis. It offers efficient implementations of various supervised and unsupervised learning algorithms, as well as tools for model selection, evaluation, and preprocessing. Scikit-learn is built on other popular Python libraries like NumPy, SciPy [27], and Matplotlib, making it a powerful and versatile tool for predictive data analysis.

In addition to its core functionalities, Scikit-learn has been instrumental in evaluating the performance of the prepared models by providing a suite of metrics. These metrics encompass diverse aspects of model performance, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Leveraging these metrics, comprehensive assessments of model effectiveness have been conducted, guiding refinement efforts and aiding in the selection of optimal models.

Moreover, Scikit-learn's utilities have facilitated the division of datasets into training, testing, and validation sets, crucial for model development and evaluation. Additionally, Scikit-learn's Recursive Feature Elimination with Cross-Validation (RFECV) [28] has been employed in conjunction with models such as Logistic Regression for feature selection. This technique iteratively prunes less informative features based on their contribution to model performance, enhancing model interpretability and efficiency. Overall, Scikit-learn's comprehensive functionalities have played a pivotal role in various stages of the project, ranging from model evaluation to feature selection, facilitating robust and insightful analysis.

Additionally, **TensorFlow** [29], a powerful open-source machine learning framework developed by Google, serves as a keystone in modern deep learning research and applications. It provides a flexible and scalable platform for building, training, and deploying machine learning models across a variety of domains. TensorFlow offers extensive support for neural networks, including traditional feedforward networks, recurrent neural networks (RNNs), convolutional neural networks (CNNs), and more advanced architectures like transformers.

Within the scope of this thesis project, TensorFlow has been predominantly utilised in conjunction with the high-level **API Keras** [30] to construct various Neural Network mod-

els. Keras simplifies the process of building neural networks by providing a user-friendly interface while leveraging the computational power and efficiency of TensorFlow as its backend. This facilitated the creation of diverse models tailored to the specific requirements of the project, including both traditional fully connected networks, often referred to as Multi-Layer Perceptron (MLP) and specialised architectures such as convolutional neural networks (CNNs).

By harnessing the capabilities of TensorFlow and Keras, the project was able to explore and experiment with a wide range of neural network architectures. This integration of TensorFlow and Keras not only streamlined the development process but also ensured performance, making it an invaluable asset in advancing the objectives of the project.

2.1.2.1 Feature Extraction and Dimensiontality Reduction

Feature selection and extraction is of paramount importance in machine learning (ML), since it facilitates the simplification of models, reduces the computational complexity of the algorithms employed, and improves generalisation of the models. There are many techniques that are dedicated to it, such as filter methods that, as it name suggests, involve the application of a filter to a set of features; wrapper methods which assess the performance of feature subsets in order to inform the selection process (forward selection, backward elimination, and Recursive Feature Elimination (RFE)) and; embedded methods which integrate within the model training process such as LASSO (Least Absolute Shrinkage and Selection Operator). The choice of feature selection technique depends on the size of the dataset, the dimensionality of the features and, the computational resources available.

In this project, **Recursive Feature Elimination with Cross Validation (RFECV)** is employed as a systematic method for selecting the most informative features for a model. As previously mentioned, it is a wrapped method that reduces overfitting by removing less relevant features, thereby improving generalisation and model efficiency.

Combining RFECV with LR, optimises feature selection, ensuring the model focuses on the most discriminative features for accurate predictions.

On the other hand, **PolynomialCountSketch** [31] is also employed as a technique of dimension reduction and characteristic extraction, enabling the approximation of polynomial expansions of input functions. This technique transforms the characteristics to a higher characteristic space, thereby enabling the capture of non-linear relationships in a simpler manner.

One limitation of PolynomialCountSketch is that the polynomial expansion may result in an increase in the combination of number of features, which can complicate computation and increase the probability of overfitting. Nevertheless, an attempt is made to address this issue by employing the "Count Sketch" technique, which reduces the dimensionality of the expansion by projecting the characteristics into a space of reduced dimension while maintaining the original polynomial relations [32].

Finally, the **Non-negative Matrix Factorization (NMF)** [33] was employed as the additional part of the line of work. This is a matrix decomposition technique that is employed to reduce the dimensionality of data. The operation of this unsupervised learning algorithm is based on factoring a non-negative matrix into two lower-range matrices whose product approaches an optimal solution of the starting matrix [34].

It is frequently employed in the context of ambiguous data sets and low predictive power, yet it preserves the original structure of the data. Moreover, it is employed as a preliminary step in the process of data preparation and classification. Nonetheless, the primary limitation of this technique is its computational cost, which is due to the necessity of iterative algorithms for its execution.

2.1.2.2 Classifiers

The field of machine learning is divided into three principal types of model approach, supervised learning, unsupervised learning and reinforcement learning. The models differ in the manner in which they are trained and the quality of the training data required.

Supervised learning represents a cornerstone in machine learning, characterised by its reliance on labelled datasets for training, usually done bay a data scientist before being used. These algorithms aim at learning the mapping between input data and corresponding output labels. By iteratively adjusting internal parameters to minimise the disparity between predicted and true labels, supervised learning models strive to generalise patterns inherent in the data. Once the model has learned the patterns between input and output data, it is able to classify new and unseen information to predict insights [35].

In contrast, unsupervised learning eschews the luxury of labelled data, instead tasked with uncovering structures within unlabeled datasets, in other words, these models are trained on raw and unlabelled training data. This autonomous exploration of data characteristics is facilitated through clustering algorithms into a specific number of groups and dimensionality reduction techniques. Furthermore, this approach is frequently employed during the initial exploratory phase with the objective of more fully comprehending the datasets.

In opposition to aforementioned, the algorithms learn to react to an environment autonomously. Reinforcement learning is characterised by the interaction of an agent with an environment to achieve a specific goal. Unlike supervised and unsupervised learning, reinforcement learning operates in an environment where explicit feedback in the form of rewards or penalties guides the learning process. Through a series of sequential decisions, the agent learns to maximise cumulative rewards by exploring the environment and exploiting learned policies. [36]

In summary, the distinction between supervised, unsupervised, and reinforcement learning methodologies extends beyond their training data requirements to encompass their utility and suitability within diverse applications. Supervised learning is particularly suited to scenarios where explicit predictions or classifications are paramount. In contrast, unsupervised learning is a valuable tool for exploratory data analysis and preprocessing tasks. Reinforcement learning, with its focus on sequential decision-making, offers unique advantages in dynamic environments where optimal strategies evolve over time.

In this project, to develop a method for detecting metal frailty, which is to say, for classifying individuals as frail or not frail. In order to achieve this objective, an algorithm based on supervised machine learning will be employed, utilising classifiers that play a pivotal role.

In the field of machine learning, a classifier is an algorithm that is capable of automatically sorting or categorising data into one or more predefined classes. The terms “targets”,“labels”, and ”categories” are all used to describe classes, in this study, the class is being frail (1) or not (0).

The classifiers used during the work, implemented with the aforementioned Scikit-learn and TensorFlow libraries, are defined below.

First, Logistic Regression (LR) [37] is a binary classification method predicting the probability of an instance belonging to a class.

On the other hand, Neural Networks are comprised of interconnected layers of numerous

components. The network components include neurons, connections, weights, biases, propagation functions and a learning rule [38]. The combination of these elements is particularly adept at learning complex patterns and relations from data, and is able to adapt to changing circumstances. Neurons receive inputs and, due to the involvement of connections, which include weights and biases, regulate the transfer of information.

The functioning of the network is based on a series of layers. The initial layer, designated the input layer, is responsible for extracting features from the input dataset. In general, each feature in the input layer is represented by a node on the network. Subsequently, the data is passed through the hidden layers, multiplied by the corresponding weights, added together, and then passed through an activation function. During the training phase, neural networks adjust their hyperparameters based on input data in order to make predictions or decisions. They are frequently employed in the processing of large-scale and high-dimensional data.

In the case of a binary classification problem, as is the case with the project in question, the output layer employs a binary activation function, which predicts whether the input data belongs to the true class (frail) or not. As the network iteratively refines its hyperparameters, it adapts to classify the different subjects. The objective is to emulate the functionality of the human brain.

In the context of supervised learning, the network generates outputs based on inputs without consideration of the surrounding context. This results in discrepancies between the generated outputs and the desired outputs, which are referred to as errors. The objective is therefore to reduce these errors by changing the parameters iteratively until an acceptable level of performance is reached.

A total of seven types of neural network (NN) were identified [39], with the Perceptron being the most basic and straightforward architecture, it consits of one neuron that applies the activation funtion obtaininga binary output without hidden layers. In the next network, Feedforward Network (FFN) data flows in a unidirectional manner from the input to the output through connected hidden layers, although they might not be necessarily present in the newtork. The Multilayer Perceptron (MLP) is a type of FFN that incorporates multiple hidden layers and activation functions, however, this network is bi-directional (forward and backward propagation). The next stage of development was the introduction of the Convolutional Neural Network (CNN) specifically tailored for processing grid-like data such as images. CNNs preserve the spatial structure of the input data, enabling them to capture intricate patterns and features. It was followed by the Recurrent Neural Network (RNN). The latter was designed for the processing of sequential data, utilising feedback loops. Then, a different strategy that predict is applied in Radial Basis Networks (RBN), which consists

of a layer with neurons employing the Radial Function as an activation function. Finally, the Long Short-Term Memory (LSTM) was developed to overcome the vanishing gradient problem of the previous architecture.

Neural networks exhibit a number of significant advantages, including their capacity to learn complex patterns and make accurate predictions across a wide variety of domains, from image recognition to natural language processing. Their adaptable architecture and capacity to automatically adjust their hyperparameters render them highly versatile and capable of addressing a multitude of problems [40, 41] .

Nevertheless, neural networks also present significant disadvantages. Training neural networks necessitates the availability of a substantial quantity of data and computational resources, which can be costly in terms of both time and resources, but in health is not often a problem. Furthermore, the opacity inherent in their internal functioning can make it challenging to interpret their decisions, which may limit their applicability in contexts where transparency and explicability are fundamental, such as in medical or legal applications. Additionally, neural networks may be susceptible to overadjustment, whereby they may learn irrelevant or noisy patterns of training data, which may affect their performance on unseen data.

Furthermore, within the supervised learning classifiers employed, the **Random Forest Classifier** [42] was utilised. The operation of the Random Forest Classifier is based on the concept of decision trees, whereby each node represents a feature and each branch represents a test result performed on that feature. This classifier consists of independent decision trees based on the combined predictions of each tree serving as the basis for the final result. Its widespread popularity can be attributed to its adaptability, which enables it to handle complex datasets and avoid overfitting. [43]

As the final classifier, the **XGBoost** (Extreme Gradient Boosting) [44] algorithm was employed. It represents an advanced implementation of the gradient boosting machine learning algorithm. This classifier is typically employed in the context of classification problems, as is the case of this project. The operation of the classifier is based on the principle of building a model in a stage-wise manner, whereby predictors are sequentially added to correct the errors made by the existing ensemble. The fundamental principle of gradient boosting is to optimise a loss function by combining weak learners, typically decision trees, to create a robust predictive model. Furthermore, the use of regularisation methods, such as L1 or L2, is recommended in order to avoid overfitting and to improve the generalisability of the model.

2.1.3 Hyperparameters Search

In relation to hyperparameters tuning, in other words, the process of selecting the optimal values for a machine learning model's hyperparameters. Since neural networks have many hyperparameters to adjust, such as the number of layers, the learning rate, epochs..., different strategies are used to find which of them are the most optimal. These approaches aid to improve model performance, reduce overfitting and underfitting, optimize resource utilization and enhance model generalizability and interpretability.

Optuna [45] is a cutting-edge hyperparameter optimization framework, it automates the process of tuning hyperparameters for machine learning models using different techniques. Optuna will be use to optimize hyperparameters for previous developed NN and CNN built with TensorFlow and Keras, enhancing their performance and efficiency.

The **GridSearch** [46] algorithm was also used as a technique to adjust hyperparameters. This technique is based on searching for each combination of parameters specified in a grid, those most optimal parameters that achieve the best performance for a specific metric. In this way, it automates and facilitates the search process for hyperparameters

2.2 Related work

In this section, existing literature is examined by analyzing a range of studies. It is aimed to identify trends, assess methodologies, and pinpoint gaps in knowledge.

2.2.1 Frailty definitions

In the corpus of analysed research, the first striking element is the lack of a common clinical definition when conducting a study. Among the various interpretations, the prevailing definition of frailty is the one introduced by Fried et al. [47], typically referred to as Fried Frailty Phenotype (FFP). Based on this phenotype, frailty has been delineated as a clinical syndrome characterised by the presence of three or more of the following criteria: inadvertent weight loss, self-reported exhaustion, diminished grip strenght, reduced walking speed, and decreased physical activity. Furthermore, it identifies an intermediate stage, pre-frailty, indicative of individuals at elevated risk of developing frailty.

The second most prevalent definition of frailty entails the utilization of a Frailty Index (FI) [48], proposed by Rockwood et al.[49], this method quantifies the proportion of deficits evident in a patient, with each deficit being assigned a score based on its severity

or frequency.

As it is dependent upon clinical judgment of the interpretation of results from history-taking and clinical examination, it appears to be time-consuming, therefore, the Clinical Frailty Scale (CFS) emerge as a more accessible alternative assessing an individual's level of frailty by considering aspects such as comorbidity, cognitive impairment, and disability. Nevertheless, FI is more effective in correlating frailty with mortality, but despite its predictive validity, it is not widely used clinically due to its practical concerns.

In addition, other studies have developed definitions of frailty based on specific variables characteristic resulted from functional decline and cognitive impairment, such as hospitalisation, mortality [50], gait speed or grip strength. The aforementioned elements collectively contribute to the enhancement of heightened sensitivity towards minor stressor events.[51, 52]

Finally, some researchers have integrated the detection of cognitive frailty into their investigations, applying different metrics mentioned before such as FFP or CSF. In some cases, the procedure was conducted in a manner that was contrary to the standard procedure, the detection of mental frailty has been used to obtain the physical frailty of patients. Furthermore, in Taiwan another dimension, social frailty [53], has been identified using demographic and physical data.

2.2.2 Data sources

The majority of studies analysed make use of Machine Learning to obtain results. In order to develop the models that are subsequently trained, they require a series of input data. For the purpose of application in mental fragility, either Electronic Health Record (EHR) or data obtained from gait and physical activity are typically employed.

In the context of EHR, the data in question originate from a variety of sources, including hospitals, nursing homes, and age surveys conducted in numerous countries. The article by Gómez-Caberro et al. [54] provides an excellent example of how data sets can be collected from different cohorts such as four different European populations that were previously analysed in their respective studies of adults over 65 years. Furthermore, it is important to highlight that the study employs machine learning as a technological proposal for identifying frailty with FFP serving as the criteria to be followed in order to achieve its objective. On the other hand, in this instance, the data originate from administrative records compiled for a number of individuals residing in nursing homes in Queensland, Australia [55]. Nevertheless, in this instance, the IF is employed for the purpose of identifying frailty.

In contrast, physical activity data are typically gathered through sensors on the wrist, pendant, legs, and other parts of the body, depending on the intended use [56].

It is notable that the majority of studies utilise databases derived from previous longitudinal age-studies, with a limited number of studies conducting their own data collection. This is a consequence of the extensive variability of databases that are currently available, which serves to avoid repetitions and to enhance the generalisability of the results proposed by each study.

2.2.3 Machine Learning applied to frailty modeling

The types of Machine Learning models applied to frailty can be classified into three main categories.

- **Prediction**. The objective of prediction models is to forecast the likelihood of an event occurring in the future. This category encompasses models developed using temporal data with the objective of predicting the future onset of frailty.
- **Detection**. The second category is that of detection. In this context, the binary classification problem is addressed, with the target classes being "frail" and "non-frail". This type of model is designed to identify the presence or absence of frailty in an individual at a given point in time.
- **Classification**. This entails a multi-class classification problem, with the target classes being "frail", "pre-frail" and "non-frail". However, additional categories may also be included. This category is less common because patients are often grouped into "robust" and "non-robust" categories in order to compensate for data imbalances.

The algorithms employed in the studies analysed are typically those that are most frequently and commonly used, including LR, SVM and RF [57, 58, 59, 60]. Consequently, several of these algorithms have been used in the design of the subsequent models, which are specified in Chapter 3. Furthermore, in combination with the aforementioned models, other techniques are employed. For instance, Hassler et al. [57] performed a feature selection process using the *Boruta* algorithm with an RF wrapper method prior to training.

Conversely, convolutional neural networks (CNNs) have also been employed, but were applied directly to images [61]. Nevertheless, in this project the data has been configured in such a way that they could be applied to a CNN , as will be discussed in more detail in Section 3. Finally, it should be noted that some studies have applied Deep Learning to enhance the performance of simpler models [62] .

CHAPTER 3

Architecture

3.1 Introduction

In this chapter, we cover the design phase of this project, as well as implementation details involving its architecture. Firstly, we present an overview of the project, divided into several modules. This is intended to offer the reader a general view of this project architecture. After that, we cover in depth each module separately.

3.2 General View

The project has been structured into two distinct phases. The initial phase involved the analysis of input data from 5303 participants, with 6583 features identified. Given the substantial number of features, the primary objective of this stage was to identify the most optimal set for subsequent analysis by driving a a feature reduction process. For this purpose, the Recursive Feature Elimination with Cross-Validation (RFECV) algorithm was used as a selection method.

Following this, the project will employ Neural Networks to explore various architectures derived from the data obtained in the initial phase. This stage will leverage the previously reduced feature set to identify the optimal design. Both traditional Neural Networks and Convolutional Networks will be utilised, aided by a hyperparameter search engine to optimize model performance. The aforementioned structural configuration has been depicted in the Illustration 3.1.

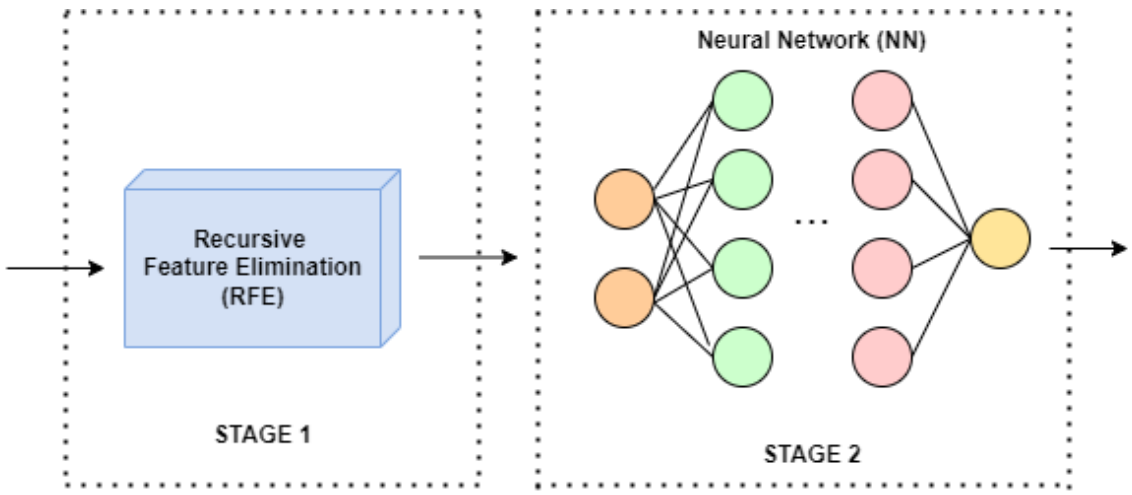

Figure 3.1: Project architecture

3.3 First stage: Recursive Feature Elimination

The selection of features is a critical step in the modeling process, as it holds substantial sway over the model's performance. A number of techniques can be employed in order to identify the most suitable features, one common method is the Recursive Feature Elimination (RFE).

RFE is a feature selection technique widely employed in machine learning to identify the most relevant features within a dataset. RFE operates iteratively by recursively removing features and assessing their impact on model performance, ultimately selecting the subset of features that maximizes predictive accuracy. In this project, RFE serves as a pivotal component in streamlining the feature space, essential for enhancing model performance.

The principal motivation for integrating RFE into this project stems from the need to mitigate the curse of dimensionality inherent in datasets with a large number of features. By systematically pruning irrelevant or redundant features, RFE not only reduces computational complexity but also improves model efficiency and robustness. In essence, RFE facilitates a more focused analysis by pinpointing the subset of features crucial for accurate

predictions, thereby enhancing model performance.

In the implementation of this algorithm, two main parameters must be selected: firstly, the number of features to be chosen, and secondly, the estimator to be used to assist in the selection of those features.

Additionally, Logistic Regression (LR) is chosen as the estimator used in conjunction with RFE for several reasons. Logistic Regression is well-suited for binary classification tasks, in alignment with the project's objective of detecting frailty. Furthermore, LR operates efficiently even with limited data, making it suitable for scenarios with a moderate sample size like the one in this project. By coupling RFE with Logistic Regression, the aim is to harness the interpretative power of both techniques in a combined manner, enabling the identification of key predictors while maintaining model simplicity and performance.

The decision to employ Logistic Regression in tandem with RFE over other models is mainly due to Logistic Regression's simplicity which allows for easier implementation, interpretation, and analysis of results.

The RFE was implemented using the Scikit-learn Python library from its Feature Selection module [63], usually employed for the purpose of feature selection and dimensionality reduction on sample sets in order to enhance the accuracy scores of estimators or to enhance their performance on very high-dimensional datasets.

In order to define the algorithm, the aforementioned hyperparameters were specified. Logistic Regression was the estimator employed. It utilised a type of regularisation, namely Ridge regularisation (L2), which penalises high coefficients by adding their squared magnitude to the cost function. This helps to prevent overfitting and improves the ability to generalise. Additionally, the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimisation algorithm was defined, which approximates the inverse of the Hessian matrix required for optimisation without storing the entire matrix in memory.

Conversely, once the estimator has been defined, the RFE combined with cross-validation (RFECV) is implemented. It performs RFE in a cross-validation loop. In this context, five-fold cross-validation is specified, and the hyperparameter that specifies the minimum number of features to be selected.This is the point at which the feature selection process will cease, in this case, when the specified threshold is reached, set at 100 features.

Finally, in order to visualise the results, the number of characteristics chosen is plotted against the CV Score using the Matplotlib and Seaborn libraries.

3.4 Second stage

3.4.1 Multilayer Perceptron

The Multilayer Perceptron (MLP) stands as a foundational architecture within the realm of artificial neural networks, characterised by its layered structure of interconnected neurons. This model has garnered considerable attention across diverse domains, owing to its capacity to discern complex patterns from data. Within the healthcare domain, MLPs have emerged as instrumental tools for predictive modeling and pattern recognition, particularly in endeavors aimed at early detection and preventative intervention.

The adoption of MLPs in healthcare settings, particularly in projects targeting early detection and prevention of health-related conditions, is underpinned by several key advantages. Firstly, MLPs excel in handling high-dimensional and heterogeneous data, making them well-suited for tasks that involve integrating diverse data sources. Given the multifaceted nature of health-related conditions, the ability of MLPs to discern complex patterns and relationships within these data streams is extremely useful.

In this context, the dataset comprises information from a substantial number of patients and a reduced set of features obtained through prior feature selection, aligns well with the strengths of MLPs. Therefore, Multilayer Perceptron can effectively leverage this rich dataset to uncover subtle associations between various risk factors and indicators of mental frailty, thereby facilitating accurate treatment.

Furthermore, the flexibility of MLP architectures allows for the exploration of intricate data dynamics and the extraction of meaningful insights from the data. Given the high variability exhibited by the dataset used, it is necessary to employ a more streamlined and consistent representation. By applying MLPs, the project aims to capture the nuanced interplay between demographic factors, medical histories, cognitive assessments, and social interactions, which collectively contribute to the manifestation of mental frailty in elderly individuals.

To maximize the performance of the MLP model, a hyperparameter optimization method has been employed, implemented through the Optuna framework. This approach enables the automated search for optimal hyperparameters, such as the number of hidden layers, the number of units in each layer, learning rate, dropout rates, and regularization strengths, thereby, streamlining the model tuning process and enhancing predictive accuracy.

A function was developed to implement the Multilayer Perceptron (MLP) using the Keras library in Python. This function defines a model in a sequential manner, which is used to design a stack of layers for a neural network. Additionally, an hyperparameter search was conducted to determine the optimal number of layers, ranging from one to five layers. Figure 3.2 shows the architectural scheme using MLPs.

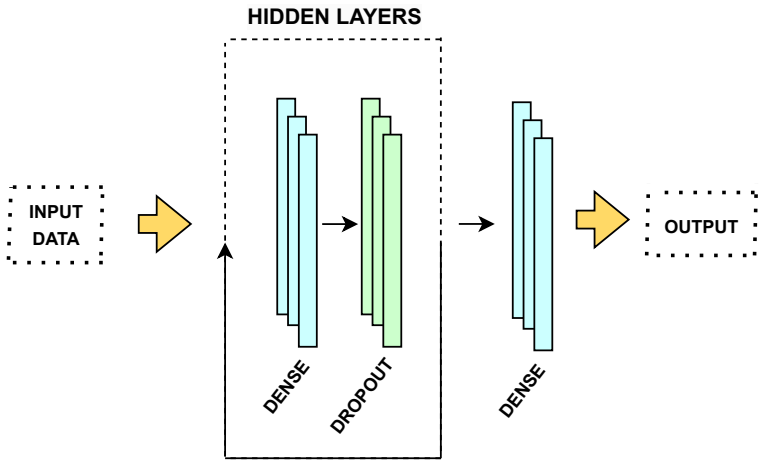

Figure 3.2: Schematic representation of Multilayer Perceptron

A fixed input layer was created to accommodate the input data. It comprised a Dense layer, a component that is widely employed in this type of neural network. In order to prevent overfitting, the previous layer was employed in conjunction with a Dropout layer. The parameter of the latter was determined by the search method.

The combination of layers was created for each layer defined by the hyperparameter search for the optimal number of layers obtained. With regard to Dense parameters, a linear transformation is performed on the input data. In particular, for each neuron specified by the search parameter, a weighted combination of all outputs from the previous layer is computed, and a bias is added. This component is defined by the number of neurons and uses a rectified linear unit (ReLU) activation function, which introduces non-linearity into the model, allowing it to learn more complex and non-linear patterns. The weights of this layer are initialised using the "random normal" kernel initialiser which draws from from a normal or Gaussian distribution. Finally, an L2-type kernel regulariser is employed, with the learning function also optimised through the search method.

Once the previous function had been defined, the model was constructed using Optuna and subsequently trained on the training subset, with validation performed on the validation

subset. The aforementioned process was executed over 50 epochs with a batch size of 128, although these values were subject to adjustment as necessary. The search parameters in Optuna were configured with the objective of maximizing one of the model's metrics, specifically accuracy. In general, 50 to 100 optimisation trials were conducted, with a final comprehensive test involving 1000 trials.

Upon completion of the parameter search using Optuna, the optimal parameters for maximizing the model's accuracy were identified. Subsequently, the model was constructed based on the aforestated optimised parameters, after which an evaluation was conducted to determine the final metrics.

3.4.2 Convolutional Neural Network

In the realm of Artificial Intelligence and Machine Learning, Convolutional Neural Networks (CNNs) have emerged as a powerful tool for various tasks, particularly those involving image and signal processing. Despite the fact this project focuses on the detection and prediction of mental frailty, leveraging the capabilities of CNNs it processes temporal health data effectively [64]. Unlike traditional Multilayer Perceptrons (MLPs), CNNs offer advantages that make them particularly suited for this application, especially when considering temporal evolution in health data [65].

CNNs are a class of Deep Neural Networks that are particularly adept at processing grid-like data structures, such as images, nevertheless, the input data from the dataset is structured in one dimension. The fundamental building block of a CNN is the convolutional layer, which applies a set of filters with convolutional operations to the input data to produce feature maps. These filters are learned during the training process and are capable of detecting local patterns. The ability of CNNs to capture spatial hierarchies in data makes them extremely powerful not only for image classification tasks, but also, to any kind of structured data, including temporal health records.

In this work, Convolutional Neural Networks (CNNs) are employed in two distinct ways. Initially, they are utilised in a manner analogous to MLPs, whereby a model is defined and applied to perform detection and prediction tasks. Secondly, CNNs are integrated with the temporal aspect of the data, allowing for the merging of different time points and the prediction of outcomes based on this combined data.

The primary motivation for using CNNs in this context stems from their ability to capture complex patterns and temporal dependencies in data, which are crucial for accurate prediction and detection of mental health conditions. CNNs excel at automatic feature

extraction from raw data. This capability is particularly beneficial in health data analysis, where important features may not be immediately apparent and can be spread across various dimensions. By designing the CNN to process data from multiple time points, the network can learn temporal dependencies and transitions over time. This temporal consideration is essential for accurately predicting the evolution of mental frailty, something that traditional MLPs might struggle with due to their lack of inherent temporal modeling capabilities.

Moreover, CNNs are known for their robustness in handling large and complex datasets. In the context of mental health prediction as frailty prediction, this robustness translates to more reliable and generalizable models that can handle the variability and noise inherent in health data. The hierarchical nature of CNNs allows for learning at multiple levels of abstraction capturing both low-level patterns and high-level trends simultaneously [66].

In a manner analogous to the MLP elucidated in the preceding section, a function is formulated to generate a model comprising convolutional layers. Prior to its definition, an adjustment must be made to reshape the data, as convolutional layers 1D require input data with three dimensions in the form of [batch_size, time_steps, input_dimension] [67]. Nevertheless, the data in question do not present this form. In order to achieve this, the characteristics are converted to timesteps. This process is applied to both the training set and the validation and test set.

Subsequently, the function that creates the model is defined in the same manner as previously supported. In a sequential manner, a model is defined to include an input layer that processes the input data. In this instance, an optimal parameters search algorithm will also be employed. The fixed input layer consists of a one-dimensional convolutional layer, with parameters such as the number of filters and kernel size determined by the optimization process. Additionally, the ReLU function is employed as an activation function. In conjunction with this layer, to reduce dimensionality, a MaxPooling layer is incorporated, which helps in the extraction of characteristics in one-dimensional data, in reducing computational load. Furthermore, a Dropout layer is added to further mitigate overfitting.

The final input set is processed by Optuna, which incorporates a BatchNormalization layer. The objective of this approach is to stabilise the training process and reduce the variability of activations. This process facilitates the convergence of the model and its stability. The structure of this network can be observed in Figure 3.3.

Subsequently, each layer introduced by Optuna includes the same components as the fixed first input layer. Once this process is completed for each of the hidden layers, a Flattening layer is added to transform the multidimensional data into a one-dimensional vector. This is followed by a fully connected layer (Dense), another Dropout and finally a

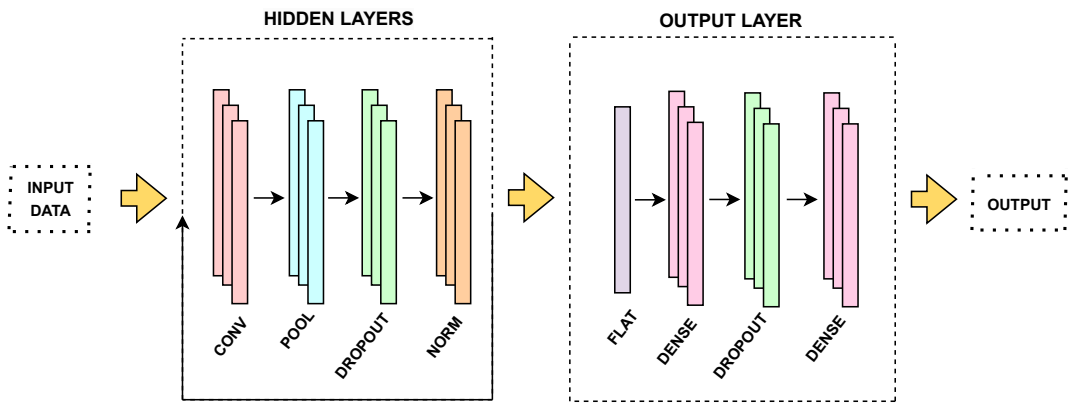

Figure 3.3: Schematic representation of the second stage architecture utilising convolutional neural networks (CNNs)

closing fully conected layer with a sigmoid activation function, which is typically used in binary classification problems. Finally, the model is then compiled using a loss function appropriate for binary classification problems, such as binary cross-entropy. This setup is subsequently employed in the search for optimal hyperparameters, which yield enhanced model performance.

In this project, CNNs were employed not only to fulfill the same roles as MLPs but also to explore their potential in integrating and analyzing data over time. The primary goal was to enhance the predictive accuracy of mental frailty by incorporating the temporal transitions. The CNN architecture was designed to take into account the temporal gaps and transitions, allowing the network to learn from both time points simultaneously. Convolutional filters were designed to capture temporal patterns and dependencies, providing a more nuanced understanding of the evolution of mental frailty. The CNN model was trained and validated on the training subset, and subsequently tested on the validation subset to ensure its effectiveness in handling temporal data.

Temporal Neural Network (TNN) design followed the same procedure for the prediction and detection of frailty, with the exception of the temporal mixed input data and the results of the metrics obtained.

3.5 Additional Stage

In addition to the primary task, a subsequent step was devised with the objective of enhancing the performance of the model. In this stage, the dimensionality of the characteristics was once again reduced, and subsequent training was conducted in two classifiers, namely the Random Forest and the XGBoost.

In order to achieve this objective, a number of tools were employed, including the construction of a pipeline and the implementation of a hyperparameter search technique, such as the GridSearch algorithm. This involved defining a grid with the hyperparameters to be adjusted. The pipeline comprises a series of steps, culminating in the training of the classifier as illustrated in the Figure 3.4.

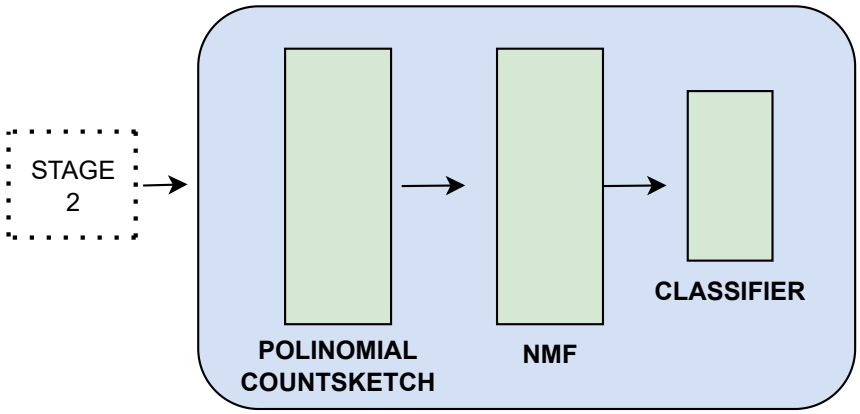

Figure 3.4: Pipeline Schematic Representation

The initial stage of the pipeline involves the application of a Polynomial CountSketch transformation to the input data. The objective of this stage is to facilitate additional interaction between the features of the input data. Furthermore, the number of components that will be produced by this transform is defined as ten times the number of features of the dataset. Polynomial CountSketch was selected for this purpose, as it avoids the iterative process of multiple training of the RFE and facilitates the discovery of complex nonlinear relationships between features that had not previously been identified.

Subsequently, a dimensional reduction is conducted utilising the Non-negative Matrix Factorization (NMF) technique with the objective of extracting intricate relationships between features. Furthermore, the number of components is automatically configured, whereby the algorithm determines the number of components according to the number of

input features. Nevertheless, this component of the pipeline presented difficulties due to the computational cost required to perform it. Consequently, the results presented in the next section will not include such a reduction.

Finally, the pipeline includes a classifier that is to be trained. Both the Random Forest classifier and the XGBClassifier were employed. Both were configured to utilise parallel processing in order to enhance computational efficiency. Nevertheless, despite the XGB utilising the GPU for enhanced speed, the RF demonstrated superior efficiency and outcomes.

Once the pipeline has been defined, hyperparameter optimization commences utilising GridSearch from the Scikit-learn library. The objective of this technique is to identify the most optimal combination of hyperparameters that maximises model performance. The hyperparameters to be optimised are specified by a separate grid, which defines the range within which the algorithm will search for each parameter. The parameters defined were exclusively for the definition of classifiers.

A cross-validation strategy is employed for the search, with the model being trained and evaluated a total of ten times. In each iteration, a different validation set is used, which helps to generalise and improve the model's performance. The metrics employed for the performance evaluation were the F1-score, precision and recall, all computed in a macro-averaged manner. The F1-score was used to select the optimal model among the results. This approach ensured that the data from the different classes was balanced to some extent.

The hyperparameter GridSearch process is executed in parallel across multiple CPU cores, thereby leveraging the available computational resources for improved efficiency. Finally, a model was defined with the best hyperparameters found by the search technique, which was subsequently evaluated.

In conclusion, the objective of this additional step was to enhance the robustness of the model and improve its predictability and generalisability.

Through the judicious application of the aforementioned algorithms within these methodological paradigms, this project endeavors to forge novel pathways toward the early detection and proactive management of mental frailty among elderly populations, thereby enhancing their overall quality of life and well-being.

CHAPTER 4

Evaluation

This chapter describes the materials and methodology used, indicating the source of the data and the preprocessing that was carried out prior to their use. In addition, the outcomes of the various algorithms developed in Chapter 3 will also be presented.

4.1 Materials and methodology

Given the objective of studying individuals over the age of 65, one of the methodologies employed to obtain data on mental frailty was to draw upon longitudinal studies concerning age. The database was obtained from the English Longitudinal Study of Ageing (ELSA) [68]. The English Longitudinal Study of Ageing (ELSA) is a large-scale longitudinal study that involves individuals aged 50 and above residing in their private residences in England.

The initial approval was granted in 2000, but it was not until 2002 when the first data for the inaugural wave were collected. The mean age of this cohort was 65 years, with a range from 50 to 100 years old.

The study subjects its participants to a personal computer-assisted interview, known as CAPI, and requires them to complete a self-questionnaire every two years. Each review

is called a wave, and a total of seven waves have been obtained, with a difference of two years between each. For the purposes of this project, Waves 5 (June 2010-July 2011) and 6 (May 2012-June 2013) were utilised. However, it is important to note that in Wave 6, new patients were added to the study. The incorporation of new patients aims to preserve the representation of younger age groups as time progresses. Each new wave includes all eligible participants from the previous wave, ensuring that younger cohorts are consistently represented over the years.

In addition to common tests, in Wave 6 nurse visit was also incorporated and has yielded a plethora of physical examination and performance data, as well as biological samples for analysis. During the visit, patients are measured for parameters such as physical function, anthropometric measurements and the collection of blood samples to analyse certain biomarkers and their DNA. In general, data on household and individual demographics are obtained, including: health,physical and psychosocial, social care (just in wave 6), work and pensions, income and assets, housing, congnite function, social participation, effort and reward, expectations, walking speed and weight.

It is important to note that among the latter are the indicator parameters to detect mental frailty, as proposed by Fried. As previously stated, this definition is employed in this work as it is widely used in the medical field and is also commonly used in other studies that apply ML techniques.

4.1.1 Data preprocessing

Prior to the commencement of the initial stage of the project, a preliminary processing of the data was conducted in order to align them with the project's objectives.

This work builds upon the research conducted by Leghissa et al.[1], which is also embedded in the MIRATAR project. The data preprocessing steps have been followed in order to ensure consistency and comparability. Nevertheless, this study introduces several improvements and alternatives with the objective of enhancing the performance of their models and incorporating novel approaches to better predict mental frailty.

From this point forward, a continual comparison of the results obtained in this project with those gathered in Leghissa's article is conducted.

Initially, the data from both Waves 5 and 6 were loaded and a function with arguments for the Fried indicator parameters was defined. These included sex, height, weight, grip strength of the dominant hand, walking time, exhaustion and activity level. Based on the FFP criteria, the participants were classified as either frail, pre-frail, or non-frail. The

results will be used as labels during the project; therefore, those characteristics that have been used for the computation of the FFP will be eliminated from the raw dataset (e.g. all grip strenght tests, activity scales, exhaustion-related questions...).

In the calculation of Fried's Frailty Phenotype for each of the waves, a treatment of each of the arguments of the function was carried out. The characteristics derived from the tests conducted at ELSA were utilised, including the sex, height, the weight. In the case of grip strength, the maximum value of the three measurements taken for the dominant hand of the subject was used. In contrast, the time taken to walk was calculated using the smallest value of the two measurements taken for each gear. Conversely, to ascertain the level of exhaustion, variables which collected whether the individual perceived everything did during past week was an effort and a challenge or whether they experienced difficulty initiating movement on a regular basis. Finally, the level of activity was assessed by determining whether the subject engaged in vigorous, moderate, or mild sports or activities on a regular basis.

To ascertain whether weight loss had been observed, the patient's weight and height were used to calculate their body mass index (BMI). The number of criteria met was increased if the BMI was greater than 18.5, as illustrated in Figure 4.1. The walking time and speed were determined depending on the patient's sex. The walking speed was calculated based on the patient's height and fulfilled the FFP criterion if it was within the specified range. Additionally, the grip strength of the dominant hand is compared according to the Figure 4.1, with the BMI obtained serving as a reference point.

Finally, if the subject's activity level is greater than or equal to 10 and their exhaustion is less than or equal to 3, this will also be added to the number of criteria met. Should the total number exceed two, it will be included in the dataset as a frail label.

Wave 6 comprised a total of 10601 participants. However, only 8054 of them underwent a nurse visit, with 2751 individuals unable to have their FFP calculated because of missing data. Consequently, the dataset to be used will be the remaining 5303 patients, of whom 52.3% are non-frail, 40.1% are pre-frail and 7.6% are frail.

Criterion	Operational Definition
• Weight loss	Unintentional weight loss $\geq$ 10% since age 60 y, or BMI $<$ 18.5
• Slowness	Time to walk 4 meters at usual pace, according to height: • $\leq$0.65 m/s for height $\leq$1.59 m • $\leq$0.76 m/s for height $>$1.59 m
• Weakness	Grip strength of the dominant hand, according to BMI: • $\leq$17 kg for BMI $\leq$ 23 • $\leq$17.3 kg for 23 $<$ BMI $\leq$ 26 • $\leq$18 kg for 26 $<$ BMI $\leq$ 29 • $\leq$21 kg for BMI $>$ 29
• Exhaustion	Any of the following during the previous month: • Low usual energy level ($\leq$3)* • Felt unusually tired in last month • Felt unusually weak in past month
• Low physical activity	Self-reported, estimated energy expenditure in kcal based on participation in six activities in the Specific Activity Scale • $<$90 kcal/wk

Figure 4.1: Fried Phenotype Criteria [2]

The Wave 6 data exhibits an imbalance, necessitating the merging of those already identified as frail and those who, despite meeting fewer indicators, may potentially develop frailty (pre-frail). This approach was adopted with the ultimate objective of helping clinicians in the early prediction of the initial stages of frailty. By merging both labels, the dataset was effectively balanced, contributing to a distribution of 52.3% non-frail and 47.7% frail individuals. This balanced distribution facilitated the transformation of the problem into a binary classification task, simplifying the analysis and interpretation of results. Furthermore, during the preparation of the data, missing and constant values were substituted in certain characteristics, as well as a scaling of all of them.

The entire procedure was applied to both waves of data. It is worth mentioning that a significant portion of participants from Wave 5 were also recruited in Wave 6. Therefore, only individuals whose FFP could be computed in Wave 6 were retained for analysis. Following the completion of all preprocessing procedures, the total number of patients in Wave 5 is 5135, which will be utilised for prediction models. Due to the absence of a nurse visit in the earliest wave, the number of features differs between waves, with 6852 features in Wave 6 and 5749 in Wave 5. However, when considering only variables common to both waves, 3821 features remain.

Furthermore, the data had to be prepared for Convolutional Neural Networks taking into account the temporal gap between waves. To this end, the patients who were present in both waves were identified with the intention of examining their evolution from one to the other. Furthermore, the patients who were present in both waves were concatenated in the same data set, which will be used at a later stage in the study of the temporal behaviour (see Figure 4.2).

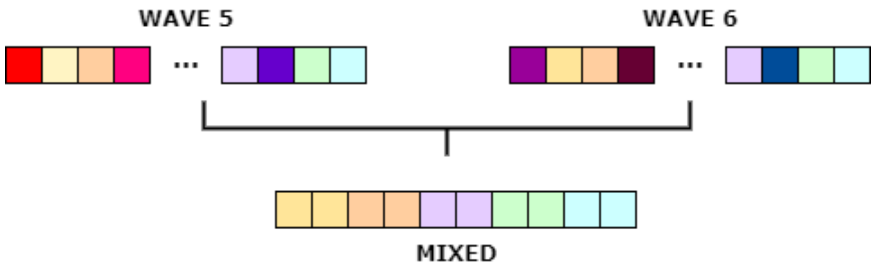

Figure 4.2: Schematic representation of *Mixed* Wave data

4.1.2 Validation and Evaluation Parameters

Once the pre-processing phase is complete, the data must be prepared in accordance with the specified training, validation and test subsets. To achieve this, Scikit-learn Selection Model library is employed to divide the prepared data into the requisite subsets. In the implementation of this process, a test size of 10% and a validation of 9% of the total data set was used and, ensuring that the division of data into the different sets maintained the same proportion of classes as in the original set.

Additionally, in the evaluation of the models described in Chapter 3, it is of the utmost importance to employ appropriate metrics that can provide a comprehensive understanding of the model's performance. Here it is outlined the key metrics employed: accuracy, precision, recall, and F1-score, explaining the rationale behind the adoption of their macro-averaged versions over the standard forms.

- **Accuracy** is a straightforward metric that measures the proportion of correctly classified instances out of the total instances. It is defined as the number of correctly identified positive (TP) and negative (TN) instances, respectively, divided by the total number of instances [69]. The simplicity of accuracy makes it a popular metric, but it may not be reliable in cases of imbalanced datasets where the number of instances in different classes varies significantly. In such circumstances, a model may achieve a high level of accuracy by simply predicting the majority class.

- **Precision** is the ratio of true positive (TP) predictions to the total number of positive predictions (TP+FP). It is defined as the proportion of instances predicted as positive that are, in fact, positive. This metric is of particular importance in applications where the cost of false positives is high. It is also known as Positive Predictive Value.

- **Recall** is the ratio of true positive (TP) predictions to the total number of actual positives (TP+FN) [70]. It is a metric that gauges the model's capacity to identify all relevant instances. It is also referred to as sensitivity or true positive rate.

- **F1-score** balances precision and recall. It is particularly useful when the dataset is imbalanced and when there is a need to find an equilibrium between precision and recall. A high F1-score indicates that a model has both high precision and high recall, making it a robust measure for evaluating model performance in complex scenarios.

Although there is not a significant imbalance in the dataset used throughout the project, it is important to prevent misinterpretations. These metrics, which are used in the case of total balancing, were macro-averaged during the evaluation. In this manner, the reformed metrics are employed to evaluate models in a more balanced way with respect to the prediction of all classes. These metrics are obtained individually for each class and subsequently, an average of of them is calculated, without giving it greater weight according to size. This ensures that the minority classes are considered in the same way as the majority classes, thus avoiding that the latter predominate in the evaluation of the performance of the model.

Furthermore, it is important to highlight that two distinct methodologies were employed to compute the metrics described above. Initially, a function was devised to calculate and return the aforementioned metrics, utilising the true labels and the labels predicted by the model. The metrics calculated by this function encompassed accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC), which served to assess the model's capacity to discern between classes. Nevertheless, this function did not take into account the possible imbalances between classes and the number of digits provided was insufficient.

To address this limitation, the previous metrics were also obtained using the function "classification_report", a detailed summary of the evaluation metrics for each classification model in the dataset. Unlike the previous function, this one also uses the actual tags and those predicted by the model, but it includes the macro and weighted average in the results. Furthermore, it was stipulated that these results were presented in four digits to facilitate comparison between the results of the various models, given that the function only yielded two digits, which limited the contrast. Finally, the F1-score is the metric that will be used to assess the performance of a model.

4.2 Results

This section will present the results obtained in the two phases of the project in accordance with the corresponding models.

4.2.1 First Stage: Recursive Feature Elimination

For each data wave, the feature selection algorithm was applied in accordance with the specifications outlined in Section 3 to select the most relevant features for the prediction model. To this end, the RFECV described above was applied, resulting in the generation of an array functioning as a mask to denote the selection status of features under consideration. This array comprises boolean elements, each corresponding to an original dataset feature. Conversely, this technique also arranges the features in an array of integers according to their relevance. The initial ranking value is presented for those features that the algorithm deems to be the most important.

In addition to the aforementioned information, a dictionary of the results of the cross-validation performed during the feature selection process can be obtained, containing performance metrics, such as the average score, for each feature configuration evaluated during the selection process. These results are useful for analysing and plotting model performance with different feature sets and, therefore, identifying the optimal combination of features that maximises model performance.

In order to facilitate the interpretation of the results, they were represented by plotting the cross-validation results obtained for each configuration of characteristics against the number of characteristics selected for each iteration of the algorithm. The following graphs illustrate the data obtained for Wave 5, Wave 6, and the combined patient data set of both waves (*Mixed*).

Figure 4.3 illustrates that at the outset of the graph, when the number of characteristics selected is relatively low, around the imposed limit of 100, the cross-validation result is relatively high. Nonetheless, the value declines rapidly as the number of additional features increases, reaching a minimum around 1000 features. Following the initial decline, the metric stabilises at approximately 1500 characteristics, where the variability of the results is reduced and oscillates around a more constant value.

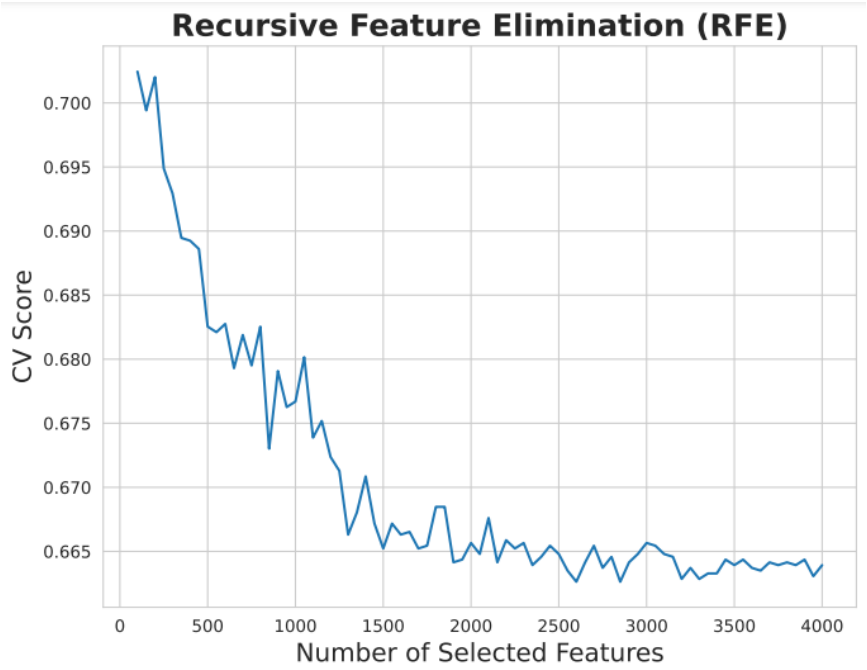

Figure 4.3: Results of Recursive Feature Elimination in Wave 5

This can be interpreted as indicating that for the initial peak, a very low number of characteristics were selected, which could be associated with an overadjustment (overfitting) to the most relevant characteristics. This could explain the initial score being so high. Nevertheless, the findings may not be generalisable.

As the number of features included in the model increases, the model may include features that are less relevant or even contribute to noise. This is why cross-validation is important. It is also possible that these new features may be reductive in certain instances.

Finally, in the final stage, which is characterised by a tendency towards stabilisation, approximately 1500 characteristics are added. At this point, cross-validation stabilises. This implies that, even if further additions are made, the performance of the model will not improve. This would be taken to indicate the optimal number of characteristics that would balance the complexity of the model with its predictive ability.

In the case of Figure 4.4, similar to the graph above, the initial behaviour begins with a high value for cross-validation when selecting a small number of characteristics, reaching a value close to 0.73, which is greater than for Wave 5. This may be indicative of the fact that the initial characteristics selected have a significant impact on the performance of the model.

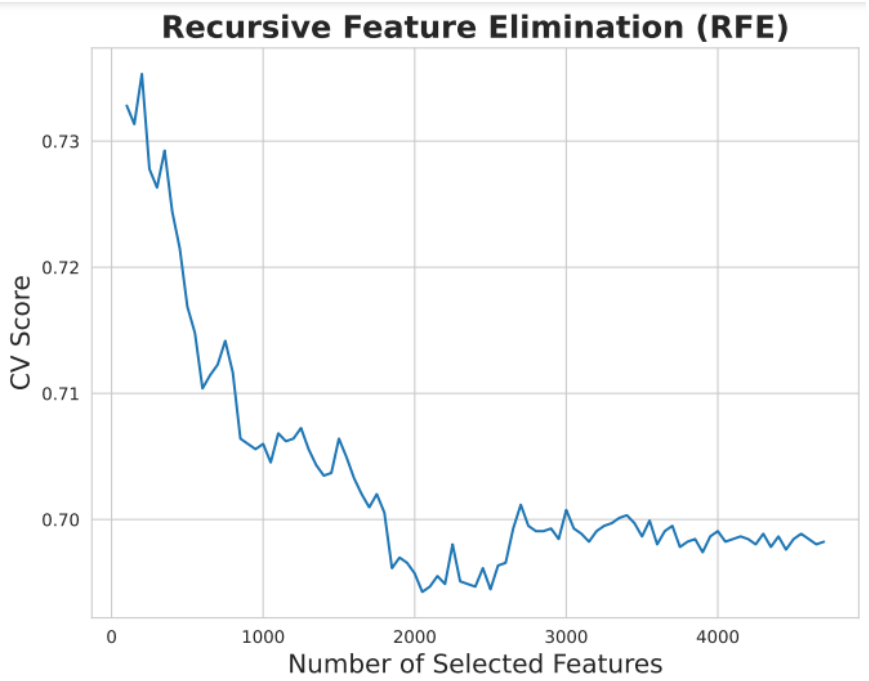

Figure 4.4: Results of Recursive Feature Elimination in Wave 6

Subsequently, there is a discernible decline in performance as the number of features increases, specifically prior to reaching 1000 features. This decline suggests that the inclusion of additional features does not contribute positively and may introduce noise. Subsequently, there is a further pronounced decline in the figure approaching 2000 characteristics, reaching a plateau where the metric score is approximately 0.70, which is comparable to the maximum score observed in the previous dataset. This would indicate that the addition of features does not significantly enhance performance.

In this instance, although there is no clear point of inflection as there was in the previous graph, a reasonable choice of characteristics could be between 2000 and 2500, where the score stabilises. Selecting fewer characteristics may result in an overfitting model or the inability to accurately capture the relationships between the data. Selecting characteristics beyond this range has been found to not significantly improve performance.

A comparison of Figure 4.3 and Figure 4.4 reveals that the differing data sets lead to different outcomes. The results obtained in Wave 6 are of a higher quality.

This discrepancy may be attributed to several factors. As the preprocessing methodology is identical for both data sets, the discrepancy cannot be due to this factor. It is possible that the difference in performance is due to the reliance on different data sets. In the case of Wave 6, the predictive capabilities or structural clarity of the data set may justify the

higher CV score.

Furthermore, the learning algorithm should not be considered a contributing factor, as the estimator employed is identical for both waves and is defined by the same hyperparameters. However, the data on which they are trained differs, with each wave representing a distinct set of data.

Another factor that may contribute to this discrepancy is the selection of distinct characteristics during the execution of the algorithm. The choice of a different set of characteristics can significantly impact the algorithm's performance. This implies that if, in the case of Wave 6, more pertinent features were selected or the selection process was conducted more optimally, this could result in superior outcomes. Finally, the most evident rationale is that the labels employed for training originate from the Wave 6 data set, which serves as a detection process. Conversely, for Wave 5, those labels were obtained from Wave 6, which serves as a prediction.

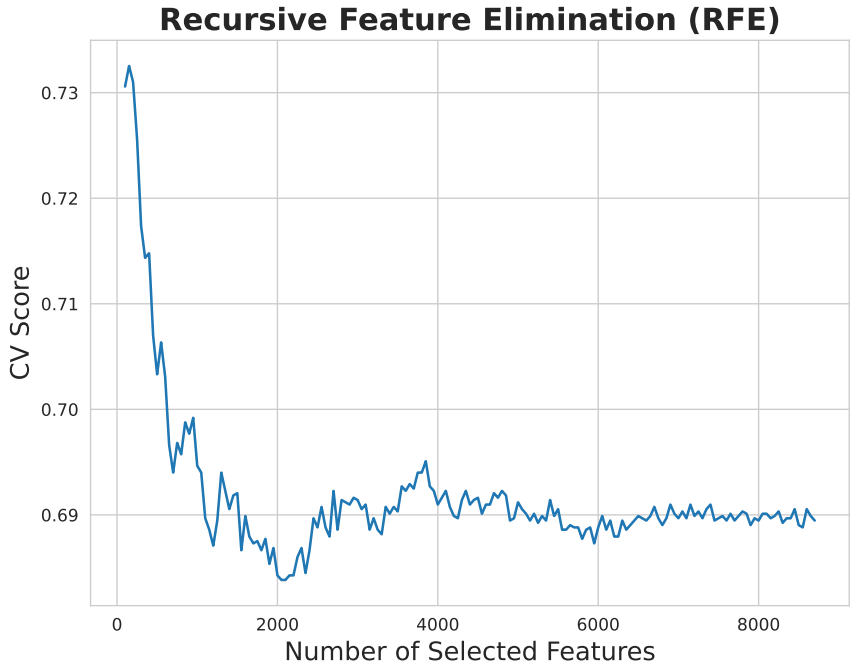

Figure 4.5: Results of Recursive Feature Elimination in *Mixed* Waves for Temporal Neural Network

Lastly, a comparison of the behaviour of the Figure 4.5 with previous situations reveals that, as with the preceding cases, when a very small number of characteristics is selected, the cross-validation score is high, reaching a value close to 0.75. This indicates that the initial features selected have a significant impact on the performance of the model. This preliminary observation suggests that these features are highly relevant.

Following the initial peak, a precipitous decline in the cross-validation score is observed as the number of characteristics is increased. This decline occurs before reaching approximately 500 features, indicating that the addition of further features does not contribute positively. From this initial decline, fluctuations emerge, suggesting that some features may be temporarily beneficial, but in general, they do not consistently enhance performance. The discrepancy in performance occurs when the number of characteristics reaches 2000, resulting in a significant decline in the score, which reaches 0.60. At this juncture, beyond the 2000 features, the score stabilizes, indicating that the addition of features does not enhance performance.

Finally, following the representation of each data set and the analysis of each graph, a total of 1500 characteristics were selected for use in the subsequent stage of the project. This was because, on average, this number of characteristics represents the point at which the performance of the model begins to increase. Furthermore, if the number of features were to be reduced even further, the representativeness and the ability of the model to capture all the variability and complexity of the problem would be compromised. The selection of fewer features may result in the loss of crucial information, which could lead to suboptimal performance and reduced model generalisation.

4.2.2 Second Stage

The results of the second stage of our project, which were previously outlined in detail in Chapter 3, are presented below.

In the pursuit to improve model performance and efficiency, optimisation strategies play a crucial role in fine-tuning hyperparameters. This includes carefully selecting ranges for parameters such as the number of units in hidden layers, which have a significant impact on the capacity and complexity of the model. Through the use of the Optuna search technique, the exploration of hyperparameters is followed with a focus on improving model performance.

4.2.2.1 MultiLayer Perceptron

In Table 4.1 are presented the results employed in Wave 5 prediction. The characteristics and differences between the various tests are detailed below.

- **Trial 1:** a total of 1000 characteristics were extracted. No L2 regularisation layer was employed, and the range of the number of units utilised in the dense layers of the

model for the hyperparameter search technique was notably broad.

- **Trial 2:** a total of 1000 characteristics were selected. The regularisation layer L2 was not employed, and the range of the number of units utilised in the Dense layers of the model for the hyperparameter search technique was expressed in powers of two, as part of an optimization strategy aimed at exploring this range thoroughly. This approach was adopted to streamline the search process and ensure a more exhaustive examination of potential configurations.
- **Trial 3:** a total of 1000 characteristics were selected. The regularisation layer L2 was employed, and the range of the number of units was expressed in powers of two for the determination of the range.
- **Trial 4:** same methodology as above, with the exception that 1500 features are utilised in this instance.
- **Trial 5:** a total of 1500 characteristics were selected. The regularisation layer L2 was employed, and the range of the number of units was expressed in powers of two for the determination of the range. In this instance, a seed was used to initialised this model.
- **Trial 6:** the same strategy as aforementioned but in a larger test of 1000 trials.

It should be noted that with the exception of the Trial 6, all trials were conducted for a duration of 50 evaluations.

NN Wave 5				
	Precision	**Accuracy**	**Recall**	**F1-score**
Trial 1	**0.74**	0.70	0.57	0.64
Trial 2	0.72	**0.72**	0.66	0.69
Trial 3	0.67	0.70	0.72	0.69
Trial 4	0.70	0.67	0.51	0.59
Trial 5	0.53	0.57	**0.76**	0.62
Trial 6	0.72	**0.72**	0.72	**0.71**

Table 4.1: Results obtained in NN Wave 5

Firstly, it can be observed that the accuracy of the tests varies considerably, with values ranging from 0.53 to 0.74. In contrast, the variability of precision remains high, with values ranging from 0.57 to 0.72. Nevertheless, there is a tendency for the precision to be around 0.70. In the case of recall, there is considerable variability (0.51-0.76), which will depend on the characteristics of the model and, therefore, make it more difficult to identify all the true

positives. Finally, the F1-score ranges from 0.59 to 0.71. These values indicate a certain balance between precision and recall.

Trial 6 may be the best value presented due to the greater number of tests conducted compared to the rest. This approach enables Optuna to test a greater number of potential combinations of hyperparameters through a more comprehensive search process, thereby increasing the probability of identifying more optimal hyperparameters that enhance the performance of the model and facilitate the accurate identification of both positive and negative instances, resulting in an elevated F1-score. Furthermore, it can be observed that Trial 6 presents more than one maximum metric result compared to other tests, corresponding to the precision and F1-score, which as previously stated are related.

Moreover, Table 4.2 presents the results employed in the Wave 6 detection. The characteristics and differences between the trails are presented in detail below.

- **Trial 1:** a total of 1500 characteristics were selected. The range of the number of units was expressed in powers of two for the determination of the range. Two hidden layers were employed.
- **Trial 2:** same strategy as before but just one hidden layer.
- **Trial 3:** a total of 1500 characteristics were selected, in this case, same layers but a regularizer was added (2 hidden layers).
- **Trial 4:** in comparison to the previous trial, a third hidden layer was incorporated.
- **Trial 5:** the same strategy as aforementioned but in a larger test of 1000 trials.

It is important to note that with the exception of Trial 6, all trials were conducted for a duration of 50 trials.

NN Wave 6				
	Precision	Accuracy	Recall	F1-score
Trial 1	0.7178	0.7137	0.7139	0.7125
Trial 2	0.7131	0.7100	0.7101	0.7090
Trial 3	0.7103	0.7006	0.7008	0.6972
Trial 4	**0.7281**	**0.7232**	**0.7233**	**0.7217**
Trial 5	0.7159	0.7100	0.7101	0.7081

Table 4.2: Results obtained in NN Wave 6

The outcomes of Figure 4.2 indicate that the results are contingent upon the specific combination of diverse hyperparameters. In the case of Trial 1, the metric values are moderately high, indicating that the hyperparameter combination is reasonably good but not fully optimal. Conversely, in the case of Trial 2, the results are slightly lower with a slight decrease in all metrics. This leads to the conclusion that decreasing the number of hidden layers does not improve the performance of the model. Similarly, Trial 3 presents the lowest values of all executions, indicating that this configuration is the least effective. Consequently, it can be concluded that the addition of regularising dense layers causes sub-optimal performance.

Nevertheless, in contrast, Trial 4 demonstrates a notable enhancement in all metrics, distinguishing itself from the other executions. It can be observed that recall (0.7233) and accuracy (0.7281) are significantly higher, indicating that the configuration is more effective. This leads to an F1-score (0.7217) that is satisfactory. Finally, the Trial 5, despite being performed in more executions (1000 trials), has not achieved better results than would be expected.

It is important to note that in Trial 5, only two hidden layers were suggested by the hyperparameter search technique. This may be the reason why the metrics are lower than in the previous trial, despite having executed it in more iterations. Therefore, it can be concluded that in order to improve the detection of Wave 6, it is highly recommended to add three hidden layers, although further investigation is required to confirm this.

4.2.2.2 Convolutional Neural Networks

This section presents the results obtained using a convolutional neural etwork model with hyperparameter optimization performed using the Optuna search technique.

Table 4.3 presents a description of the tests conducted for the purpose of Wave 5 prediction using a convolutional neural network (CNN).

- **Trial 1:** a total of 1500 features were analysed. In this instance, a hidden layer and a MaxPooling1D layer were employed. A total of 50 executions were conducted.
- **Trial 2:** Dropout was incorporated into each hidden layer for a total of 1500 features. Furthermore, a Learning Rate Schedules technique was implemented as an optimizer of the learning rate, in this case, Exponential Decay.

  It was determined that this technique should be incorporated into the model training process due to the critical nature of the learning rate as a hyperparameter in the

learning model training process. Additionally, during the manual testing conducted, it was observed that the variation in the learning rate exhibited a notable influence.

Consequently, it was determined that the incorporation of a learning rate schedule would enhance the efficacy and efficacy of the project, thereby facilitating a more rapid and stable convergence. In the case of Exponential Decay, the learning rate is reduced exponentially as the training progresses. In this instance, the subject was trained with 100 trials.

- **Trial 3:** involved training with 1500 features and 100 trials. Nevertheless, the key distinction is that the learning rate schedule varies, employing the Inverse Time Decay technique. This technique reduces the learning rate in inverse proportion to the number of iterations, maintaining a high pace in the early stages to gradually reduce it as the training progresses.
- **Trial 4:** the strategy is identical, and in this instance, Polynomial Decay is employed. This entails following a polynomial function instead of the previous approaches, which employed exponential or inverse functions.
- **Trial 5:** employs the Consine Decay learning rate schedule, which varies according to a cosine technique.

CNN Wave 5				
	Precision	Accuracy	Recall	F1-score
Trial 1	0.71	0.71	0.71	0.71
Trial 2	**0.7456**	**0.7360**	**0.7291**	**0.7289**
Trial 3	0.7307	0.7259	0.7030	0.7205
Trial 4	0.7348	0.7212	0.7130	0.7114
Trial 5	0.7408	0.7243	0.7155	0.7136

Table 4.3: Results obtained in CNN Wave 5

The results presented in Table 4.3 indicate that in the initial run, the metric values are uniformly 0.71, reflecting a combination of hyperparameters that provides an stable level of performance. Nevertheless, these values are presented with fewer decimals than those in the subsequent rows. This is because the function described in Section 3 was used, but was later discarded due to its limited comparative capacity. The function provided by machine learning library from Scikit-learn was subsequently employed, providing a detailed summary of various evaluation metrics.

Conversely, the outcomes of Trial 2 demonstrate a notable enhancement in all metrics in comparison to the preceding execution. This may be attributed to the incorporation of the

MaxPooling1D layer and ExponentialDecay, in addition to the increase in the number of trials. Trial 3, however, demonstrates a slight decline in performance compared to previous trial, as evidenced by the lower precision and F1-score values. This suggests that the combination is less optimal. The observed decrease in performance is associated with the change in the function used for optimisation of the learning rate.

Nevertheless, in Trial 4, there is an improvement in precision and recall compared to previous run, yet the results remain below those of Trial 2. Furthermore, the F1-score is lower than before, indicating that the current combination is not yet optimal. In this instance, it is necessary to rule out the optimisation function that was previously employed. Finally, Trial 5 exhibited a slight improvement compared to the previous run.

In consideration of the outcomes observed for each trial, the second trial is identified as the most effective, exhibiting the most optimal results across all evaluated metrics. This suggests that the configuration of this execution is the most effective. It can be concluded that the utilisation of the ExponentialDecay optimisation function is responsible for the generation of the most favourable outcomes. This may be attributed to the fact that this particular technique enables a gradual and continuous reduction of the learning rate.

The methodology employed to generate the data related to detection in Wave 6 presented in Table 4.4 is outlined below.

- **Trial 1:** In the initial trial, 50 trials were conducted with two hidden layers, as obtained with Optuna. The first layer consisted of 32 filters with a core size of 5, while the second layer comprised 16 filters with a core size of 5. Furthermore, the Exponential Decay technique was employed as the optimizer schedule.
- **Trial 2:** employs the same strategy as Trial 1, but with an increased learning rate in order to enhance the optimization process.
- **Trial 3:** this execution employs three hidden layers as recommended by Optuna and does not utilise any optimiser for the learning rate.
- **Trial 4:** The two-hidden-layer configuration is reintroduced, with the exclusion of a schedule optimiser and the utilisation of the rate proposed by Optuna.
- **Trial 5:** The strategy employed in the initial attempt was replicated, with the parameters of the Exponential Decay varying.

In all previous trials, the search technique was executed during 50 trials and its number was not varied. This will be taken into account in the comparison below.

CNN Wave 6				
	precision	**accuracy**	**recall**	**F1-score**
Trial 1	0.7155	0.7137	0.7138	0.7132
Trial 2	**0.7332**	**0.7232**	**0.7234**	**0.7203**
Trial 3	0.7008	0.7006	0.7006	0.7005
Trial 4	0.7246	0.7213	0.7214	**0.7203**
Trial 5	0.7173	0.7137	0.7139	0.7126

Table 4.4: Results obtained in CNN Wave 6

The results presented in Table 4.4 indicate that the values obtained for Trial 1 were moderately high, although not entirely optimal. The second run demonstrated a notable enhancement in all metrics in comparison to the previous iteration, culminating in the most favorable performance observed across all executions, since is superior to the rest in terms of all the scores that have been considered. The explanation for this improvement may be due to the use of a considerable number of layers and the use of the ExponentialDecay technique, which has already been found to improve the performance of the model, as well as improving it with respect to the previous execution. This is due to the human manipulation of the learning rate that was carried out after the observation of a slow drop of it during the evaluation.

Trial 3 exhibited the lowest values in all metrics in comparison to the other trials. In contrast to the previous cases, the introduction of a hidden layer did not result in enhanced performance. Consequently, the model was reconstructed with two hidden layers, as this configuration has been demonstrated to enhance the model's performance. In opposition to the previous trial, there was an improvement in precision and recall in Trial 4. It is possible that this is due to the removal of the learning rate optimisation resource. Nevertheless, these values remain below those observed in the second run. Nonetheless, the results demonstrate a notable distinction between recall and precision, resulting in a competitive performance. Finally, Trial 5 yielded values comparable to those observed in the initial run.

A review of the data reveals that the optimal performance was achieved in Trial 2, suggesting that the combination of two hidden layers with the ExponentialDecay technique, in conjunction with the hyperparameters identified through the search technique, has been demonstrated to yield the most optimal results for the detection of Wave 6 in Convolutional Networks.

The results of the study indicate that the convolutional neural networks (CNNs) employed in these experiments exhibited superior performance compared to the multilayer perceptrons (MLPs) previously evaluated. The CNNis particularly effective in data processing

due to its ability to capture spatial characteristics through convolutions. This capacity to extract and learn intricate spatial patterns enables CNN to surpass MLP in tasks such as this project.

The following table presents the results of the combined data from both waves (*Mixed*), as shown in Figure 4.2. These results are intended to show potential improvements by introducing temporal information. As discussed at the end of Section 3.4.2, incorporating data from multiple waves of time allows for a more comprehensive analysis by exploiting the temporal dimension, which can improve the predictive power of the model. The metric results presented in Table 4.2 highlight the impact of adding temporal information on the overall effectiveness of the model.

- **Trial 1:** The experiment was conducted in 50 iterations, utilising the hyperparameters proposed by Optune for a hidden layer comprising 16 filters and a kernel size of 4, with a relatively low learning rate (0.00001).
- **Trial 2:** This trial follows the same strategy as the previous one, but the learning rate is increased to 0.01.
- **Trial 3:** In this case, two hidden layers are used. The first layer has 16 filters with a kernel size of 2, and the second layer has 64 filters and a core size of 4. The learning rate remains the same as in the previous trial.
- **Trial 4:** employed the same strategy as in the initial trial, with the addition of the Exponential Decay learning rate optimizer.
- **Trial 5:** The same strategy as before was employed, with the addition of a second layer comprising 64 filters and a core size of 3.

In summary, all results were developed for 50 iterations. However, it should be noted that additional tests were also performed, where the number of executions was increased, but no superior results were achieved. It can be concluded that, despite the limited number of iterations, the Optuna search method is capable of identifying the most optimal parameters in a smaller number of combinations, given the data in question.

Table 4.5 presents five distinct executions, each exhibiting varying outcomes. In the initial attempt, it is possible that the low learning rate contributed to a slow convergence, resulting in a moderate performance. Additionally, the highest F1-score value was achieved in this attempt. Conversely, a slight increase in the learning rate in Trial 2 led to enhanced precision and accuracy metrics, while maintaining comparable recall and F1-score. This suggests a balance between rapid convergence and stability.

Mixed				
	precision	**accuracy**	**recall**	**F1-score**
Trial 1	0.7681	0.7665	0.7640	**0.7665**
Trial 2	0.7761	**0.7685**	**0.7641**	0.7645
Trial 3	0.7753	0.7529	0.7457	0.7438
Trial 4	0.7812	0.7549	0.7472	0.7448
Trial 5	**0.7872**	0.7588	0.7509	0.7485

Table 4.5: Results obtained in CNN mixed

In the case of Trial 3, the inclusion of a second hidden layer with more filters did not enhance the model's capacity to capture more complex features, as evidenced in a slight decline in the metrics, which could indicate an excessive adjustment in specific characteristics. The utilisation of ExponentialDecay in Trial 4 enhanced the stability of the model, resulting in high precision, and a slight increase in recall and F1-score in comparison to the previous run. Finally, the Trial 5, which incorporated a second hidden layer with additional filters, yielded the most optimal performance of its precision, achieving a value of 0.7872, the highest of all executions. Furthermore, the recall and F1-score demonstrate a satisfactory balance.

Following the comprehensive examination of the outcomes yielded by each wave and type of network, the CNN has been demonstrated to exhibit superior performance in comparison to the MLP, due to its inherent capacity to capture both spatial and temporal characteristics within the data. In the context of the combined data from the two different time waves, the CNN is able to leverage time dependencies and local characteristics more effectively. Furthermore, this aspect is of particular relevance in the context of health applications, as is the case of this project temporal patterns and spatial relationships within the data can be crucial both for diagnosis and for possible prevention or follow-up of conditions such as mental frailty.

Finally, after obtaining the previous results for each of the model alternatives, these outcomes were compared with those from the Leghissa's study, which serves as the foundation for this research as introduced in Section 4.1.1. Table 4.6 showcases the metrics for the best models from each wave, comparing their F1-scores with those reported in the original article.

	Models	**Precision**	**Accuracy**	**Recall**	**F1-score**
Wave 5	MLP	0.72	0.72	0.72	0.71
	CNN	0.7456	0.7360	0.7291	0.7289
	Leghissa	0.741	0.736	0.732	0.731
Wave 6	MLP	0.7281	0.7232	0.7233	0.7217
	CNN	0.7332	0.7232	0.7234	0.7203
	Leghissa	0.741	0.738	0.734	0.734
Mixed	CNN	**0.7681**	**0.7665**	**0.7640**	**0.7665**

Table 4.6: Comparison of the results of the project with the results described in Leghissa's article [1]

The presented findings in Table 4.6 indicate that MLP[1] networks do not emerge as a preferable alternative for improving the outcome metrics across all waves, as demostrasted in both, Leghissa's[2] study and the convolutional network analysis. However, while the F1-score value achieved in Wave 6 surpasses that of convolutional networks, the disparity in results lacks significance to definitively assert the superiority of MLP in detection tasks. Furthermore, other performance metrics fail to exhibit enhancements compared to CNNs, therefore, further exploration would be required to definitively confirm the previous conviction. Thus, it can be inferred that the design of MLPs remain questionable and might not be optimal for the proposed system, as they appear not capable in discerning relationships effectively.

Conversely, a similarity exists between the outcomes of convolutional networks and those referenced in the article concerning accuracy and recall metrics. Nevertheless, this study's model demostrates superior precision compared to the referenced article, indicating a more accurate prediction of positively identified instances. Specifically, in Wave 5, CNNs achieve a precision of 0.7456, suggesting their efficacy in predicting mental frailty. In contrast, this is not the case of Wave 6, whose precision values are lower. However, across both prediction tasks in Wave 5 and detection tasks in Wave 6, the F1-scores fail to surpass those reported in the referenced article.

Nevertheless, the most striking findings in Table 4.6 are significantly superior results obtained by incorporating data with temporal dependencies into convolutional network

[1]Wave 5 MLP results show fewer digits because they use a less powerful strategy for obtaining metric results

[2]The original results presented in Leghissa's article were obtained with a precision of three decimal places.

models. In contrast to other approaches, incorporating temporal information yields higher values across all evaluated metrics, attaining an F1-score value of 0.7665. It is evident that the performance of the model significantly surpasses all other results depicted in the comparative table.

In summary, while both MLP an CNN present certain merits in predicting mental frailty, the inclusion of temporal dependencies in convolutional network models yields markedly superior results. This underscores the importance of considering temporal dynamics in the analysis of mental frailty, ultimately enhancing predictive accuracy and informing more effective interventions.

For instance, in the context of temporal health analysis, the disparate data collected from patient records are presented as features. Nevertheless, those that are relevant may not manifest in isolation, but rather as potential patterns distributed over time. CNN is presented as a powerful tool due to its convolution and pooling capabilities, which allow for the capture of characteristics, thereby improving the capacity of the model and achieving accurate predictions.

4.2.3 Additional Stage

In order to enhance system performance, new alternatives were proposed as detailed in Section 3.5. The objective was to introduce a new stage of feature selection after the neural network, focusing on selecting the most contributive features again.

In this stage, an alternative hyperparameter search technique, GridSearch, was employed. However, the computational cost of the pipeline in conjunction with the extensive hyperparameter search, made test execution challenging.

Among all the runs performed, the maximum **F1-score** achieved was **0.7135**. Although this score matches the MLP result in predicting Wave 5, as shown in Table 4.6, the score did not yield superior results to those of the CNNs. Nevertheless, this line of research is not precluded and remains promising. The objective of reducing the number of features, as demonstrated by RFE, enhances system performance and helps the identify the most optimal contributing features. Exploring other robust models, such as SVM, would be a valuable continuation.

Consequently, access to additional resources would be beneficial for pursuing this research further. This would enable a definitive test to determine whether these ideas should be discarded or if they could introduce potentially innovations in predicting and detecting mental frailty.

CHAPTER 5

Conclusions and future work

This section presents the principal conclusions that have emerged from the design and development of this project and an outline on future work.

5.1 Conclusions

This work follows the line of research of the article developed by Leghissa [1], whose main objective aligns with that of this project: predicting early mental frailty. Compared to other studies, the development of this machine learning system has been conducted by designing various models using neural networks as the primary element. This approach contrast with other research [57, 58, 59, 60] that employs conventional and commonly used models of machine learning algorithms.

Neural networks were chosen for this study not specifically for the task at hand, but rather because of the nature of the data available. The dataset is highly variable, requiring a simplified and unified representation. Therefore, the combination of feature reduction techniques with neural networks was used to address this challenge. These algorithms are capable of capturing complex relationships and patterns in the collected data, as well as, incorporating the time dependencies of data collected at different points in time, which is

typically in patient follow-up and prevention in the health sector. Additionally, the research line developed in this project differs from Leghissa's article in the initial stage of feature extraction used in other developments, with the technology employed here standing out for its simplicity and effective results.

One of the main limitations encountered at the beginning of the development process was the difficulty in finding the most optimal hyperparameters for the models. Consequently, a search technique was incorporated to facilitate the design of the models. The performance metrics were compared using the macro-averaged F1-score to select the models with the best performance.

Additionally, in relation to the architecture developed in Section 3, it presents two main structures. First, the use of MLPs in Section 3.4 whose metrics are not high enough or do not improve upon existing metrics. In contrast, Section 3.4.2 discusses the incorporation of convolutional layers, which improves the F1-score but still do not introduce significant improvements in the study. Nevertheless, when these convolutional layers were employed to detect hidden time dependencies in the data, the resulting model achieved superior and better results than all previous ones.

In conclusion, by using CNNs, the obtained scores are superior and enhance the prediction of mental frailty in older adults. This project successfully developed a model that performs correctly, improving the predictions of previous research by capturing the temporal dependencies present in the data.

5.2 Achieved goals

This section highlights the key achievements of the project and the strengths of the system developed. During the development of the project, all the objectives were achieved and have served as a guideline for the line of research.

The first objective (1) refers to the ideal of reducing the number of tests to be carried out on patients by means of feature extraction techniques. Section 3 introduces the Recursive Feature Extraction technique, which identifies the combination of most relevant features for prediction in each wave of data. The results can be used to establish a minimal set of tests necessary for patient evaluation, thereby reducing the number of assessments a patient must undergo and alleviating the workload of clinicians by eliminating the collection of non-influential data.

To address the second objective (2), as described in the Section 3, numerous models were explored and developed to enhance the system performance. Various configurations of MLPs and CNNs were tested, with their performance optimised through hyperparameter tuning using the Optuna search technique. This approach involved the addition of hidden layers and the adjustment of hyperparameters. Additionally, a brief experiment was conducted using XGBoost and SVM as an alternative line of research.

Finally, the models described were evaluated a system whose performance could be effectively used in conjunction with diagnostic tests to predict frailty. For the dataset combining different time points, the best results were achieved by the model utilizing convolutional networks. Thus, the third objective (3) of the project was successfully met by identifying a robust model that can enhance the prediction of frailty through the integration of temporal data.

5.3 Future work and challenges

This section provides a brief overview of the principal challenges and prospective avenues for future research that may emerge from this development. The proposals for improvement and future developments are presented below:

- With regard to **classifiers**, one potential avenue for investigation would be to utilise more robust and powerful classifiers, such as XGBClassifier, Random Forest or Support Vector Machine. This approach would not only be capable of identifying complex patterns but would also be able to uncover hidden relationships, thereby yielding superior outcomes. Nevertheless, it is important to consider the increased computational cost of implementing such enhancements.

- In terms of future **applications**, it would be stimulating to utilise this technology not only for prediction, as demonstrated in this project, but also for detection. Such an approach would be advantageous, as early detection would enable the prediction of subsequent long-term signs of mental frailty, thereby reducing the risk of developing this condition to a greater extent in the future.

- With regard to the condition under consideration in this study, although in Section 1 Introduction we introduced the concept of frailty as a **multidimensional** syndrome, the approach of this project is based on the prediction of mental frailty. Nevertheless, the approach could be extended to other dimensions, such as the physical or social dimensions. The method could allow for the detection of a single condition, multiple

conditions simultaneously, or the probability of developing one or more conditions based on one of them.

- Concerning the **multimodality**, it would be beneficial to include medical tests as images in addition to clinical or administrative data that are already incorporated, in order to gather as much information as possible to make a more accurate diagnosis of frailty and to enhance the performance of the models.

- In relation to the implementation and treatment of the patient, it would be beneficial to develop systems that operate in real time and collect data from a certain age. These data could then be incorporated into a portable device, such as a mobile phone, and used as applicable data in models.

- In the context of the MIRATAR project, the system's data can be used to develop **treatments** to prevent or treat mental frailty. By identifying early signs and suggesting personalised interventions, the system helps to reduce the onset of frailty and increase the effectiveness of treatment.

Additionally, while the systems show promise, the potential computational costs and time required for further improvements are non-negligible considerations that must be addressed in future work.

While there are promising opportunities for further research and notable advantages, a significant drawback remain. As discussed in the introduction, the lack of a consensual definition of frailty within the medical community presents an ambiguity and a fundamental challenge. Without a standarised definition, the designs and models developed by any similar work, including this project, will be susceptible to variations based on new criteria. Consequently, their reliability and validity may fluctuate, posing a challenge for consistent application and interpretation.

APPENDIX A

Impact of this project

This appendix provides a quantitative and qualitative assessment of the project's potential social, economic, environmental and, ethical impacts.

A.1 Social impact

As previously discussed, mental frailty is a widespread problem among older people, resulting in an increase concern as the population ages. By developing a machine learning system for early detection of mental frailty in elderly people, this project aims to facilitate timely diagnostic and treatment. Early detection and prediction ensure older people better aging, alleviate personal distress and reduce the wider social and economic burden associated with demographic aging.

A.2 Economic impact

Early detection of mental frailty can not only prevent worse prognoses but also reduce the associated costs. While hospitalization [71] represents a significant portion of healthcare

expenses, early prediction of mental frailty can also diminish the need for other resources, such as physiotherapy, medications, specialist consultations, general practitioner visits, and diagnostic tests.

If prediction systems similar to the one proposed in this work were implemented, it would be possible to reduce costs by providing early indications for treating this condition. Additionally, it would allow for better management and optimization of medical specialties, facilitating more effective follow-up and treatment plans, as well as improving the allocation of diagnostic tests and prevent them form invasive and expensive treatments.

A.3 Environmental impact

One of the main environmental impacts is the power consumption resulting from the complexity of the machine learning algorithm. Although efforts have been made during the development of the project to consider and minimise the power consumption of the GPU, it can still consume up to 250W.

A.4 Ethical impact

The deployment of the machine learning system for the early detection of mental frailty arise several ethical considerations that must be thoroughly addressed.

First, it is essential that the system adheres to ethical and legal requirements to protect patient data, maintaining the confidentiality of medical information.

On the other hand, the system must be designed to ensure that no community or individual is exluded. This is particularly important for elderly populations, the system should be culturally sensitive, considering different cultural perspectives on mental health. By addressing these factors, the system developed should ensure that all elderly individuals have an equal opportunity to benefit from early detection.

APPENDIX B

Economic budget

This appendix details an adequate budget to bring about the project, including hardware resources, software resources and human resources.

B.1 Hardware resources

During the development of this project, the only hardware component required was the use of a computer with the following characteristics:

- **RAM:** 8 GB
- **CPU:** Intel (R) Core (TM) i5-2400 @2.10 GHz
- **GPU:** NVIDIA Titan X
- **Operating Sysmtem:** Ubuntu 22.04.3 LTS
- **Storage:** 1 TB

The estimated cost of acquiring these component would be around of 1000€ the computer and the GPU for about 3500€.

B.2 Software resources

The software tools used in this project incur no costs, as they are open-source and free from any licensing fees.

B.3 Human resources

In order to calculate the approximate cost of the human resources used to develop this project, the number of credits associated with the completion of a final thesis must be considered equivalent to 12 ECTS. Each of these credits corresponds to approximately 25-30 hours per week, so that a total of 360 hours would be required to carry out the project.

In Spain, the average salary of a recently graduated biomedical engineer is 25000€ per year, which would be approximately 11€ per hour. With this information, the approximate total cost for the hours worked by the student is 3960€.

B.4 Resources Total Cost

In summary, if all the associated costs described above are added together, the total cost of developing the project is 8460€.

Bibliography

[1] Carlos A. Iglesias Matteo Leghissa, Alvaro Carrera. Frelsa: a database for frailty in elderly people originated from elsa and validated through machine learning models. *On press*, 2024.

[2] Paulo Chaves, Richard Semba, Sean Leng, Richard Woodman, Luigi Ferrucci, Jack Guralnik, and Linda Fried. Impact of anemia and cardiovascular disease on frailty status of community-dwelling older women: The women's health and aging studies i and ii. *The journals of gerontology. Series A, Biological sciences and medical sciences*, 60:729–35, 06 2005.

[3] Organisation for Economic Co-operation and Development (OECD). Elderly population (indicator), 2024.

[4] World Health Organization. Ageing and health, 2022.

[5] Àngel Lavado, Júlia Serra-Colomer, Mateu Serra-Prat, Emili Burdoy, and Mateu Cabré. Relationship of frailty status with health resource use and healthcare costs in the population aged 65 and over in catalonia. *Eur. J. Ageing*, 20(1):20, June 2023.

[6] Sean X Leng Xujiao Chen, Genxiang Mao. Frailty syndrome: an overview. *Clinical Interventions in Aging*, 2014.

[7] Qian-Li Xue. The frailty syndrome: Definition and natural history. *Clinics in geriatric medicine*, 2011.

[8] Linda P. Fried, Catherine M. Tangen, Jeremy Walston, Anne B. Newman, Calvin Hirsch, John Gottdiener, Teresa Seeman, Russell Tracy, Willem J. Kop, Gregory Burke, and Mary Ann McBurnie. Frailty in Older Adults: Evidence for a Phenotype. *The Journals of Gerontology: Series A*, 56(3):M146–M157, 03 2001.

[9] Marie-Annick Le Pogam, Laurence Seematter-Bagnoud, Tapio Niemi, Dan Assouline, Nathan Gross, Bastien Trächsel, Valentin Rousson, Isabelle Peytremann-Bridevaux, Bernard Burnand, and Brigitte Santos-Eggimann. Development and validation of a knowledge-based score to predict fried's frailty phenotype across multiple settings using one-year hospital discharge data: The electronic frailty score. *eClinicalMedicine*, 44:101260, 2022.

[10] Felipe Diaz-Toro, Gabriela Nazar, Claudia Troncoso, Yeny Concha-Cisternas, Ana Maria Leiva-Ordoñez, Maria Adela Martinez-Sanguinetti, Solange Parra-Soto, Nicole Lasserre-Laso, Igor Cigarroa, Lorena Mardones, Jaime Vásquez-Gómez, Fanny Petermann-Rocha, Ximena Diaz-Martinez, and Carlos Celis-Morales. Frailty index as a predictor of mortality in middle-aged and older people: A prospective analysis of chilean adults. *International Journal of Environmental Research and Public Health*, 20(2), 2023.

[11] Peter Hanlon, Barbara I Nicholl, Bhautesh Dinesh Jani, Duncan Lee, Ross McQueenie, and Frances S Mair. Frailty and pre-frailty in middle-aged and older adults and its association with multimorbidity and mortality: a prospective analysis of 493 737 UK biobank participants. *Lancet Public Health*, 3(7):e323–e332, July 2018.

[12] Christopher McNeil Mina Khezrian, Phyo K. Myint and Alison D. Murray. A review of frailty syndrome and its physical, cognitive and emotional domains in the elderly. *Geriatrics (Basel, Switzerland)*, 2017.

[13] Francesca Ferrari Pellegrini Aurelio Maria De Iorio Chiara Fazio Raffaele Federici Elena Gallini Umberto La Porta Giulia Ravazzoni Maria Federica Roberti Marco Salvi Irene Zucchini Giovanna Pelà Fulvio Lauretani, Yari Longobucco and Marcello Maggio. Comprehensive model for physical and cognitive frailty: Current organization and unmet needs. *Frontiers in psychology*, 2020.

[14] John E. Morley, Bruno Vellas, G. Abellan van Kan, Stefan D. Anker, Juergen M. Bauer, Roberto Bernabei, Matteo Cesari, W.C. Chumlea, Wolfram Doehner, Jonathan Evans, Linda P. Fried, Jack M. Guralnik, Paul R. Katz, Theodore K. Malmstrom, Roger J. McCarter, Luis M. Gutierrez Robledo, Ken Rockwood, Stephan von Haehling, Maurits F. Vandewoude, and Jeremy Walston. Frailty consensus: A call to action. *Journal of the American Medical Directors Association*, 14(6):392–397, 2013.

[15] Alistair Wilkes. Early frailty can now be detected using machine learning.

[16] Francisco Anabitarte-García, Luis Reyes-González, Luis Rodríguez-Cobo, Carlos Fernández-Viadero, Silvia Somonte-Segares, Sara Díez del Valle, Eneritz Mandaluniz, Roberto García-García, and José M. López-Higuera. Early diagnosis of frailty: Technological and non-intrusive devices for clinical detection. *Ageing Research Reviews*, 70:101399, 2021.

[17] R.C. Ambagtsheer, N. Shafiabady, E. Dent, C. Seiboth, and J. Beilby. The application of artificial intelligence (ai) techniques to identify frailty within a residential aged care administrative data set. *International Journal of Medical Informatics*, 136:104094, 2020.

[18] Gestión predictiva, personalizada, preventiva y participativa (4P) de la fragilidad y la multimorbilidad para la transición digital de la economía de los cuidados (MIRATAR). `https://arcoresearch.com/portfolio-items/miratar/?portfolioCats=1397`. [Accessed 03-06-2024].

[19] Welcome to Python.org — python.org. `https://www.python.org/`. [Accessed 02-06-2024].

[20] pandas - Python Data Analysis Library — pandas.pydata.org. `https://pandas.pydata.org/`. [Accessed 20-05-2024].

[21] Matplotlib &x2014; Visualization with Python — matplotlib.org. `https://matplotlib.org/`. [Accessed 20-05-2024].

[22] seaborn: statistical data visualization &x2014; seaborn 0.13.2 documentation — seaborn.pydata.org. `https://seaborn.pydata.org/`. [Accessed 20-05-2024].

[23] NumPy - — numpy.org. https://numpy.org/. [Accessed 20-05-2024].

[24] pickle — Python object serialization — docs.python.org. https://docs.python.org/3/library/pickle.html. [Accessed 20-05-2024].

[25] Project Jupyter — jupyter.org. https://jupyter.org/. [Accessed 20-05-2024].

[26] scikit-learn: machine learning in Python &x2014; scikit-learn 1.4.2 documentation — scikit-learn.org. https://scikit-learn.org/stable/. [Accessed 20-05-2024].

[27] SciPy - — scipy.org. https://scipy.org/. [Accessed 20-05-2024].

[28] sklearn.feature_selection.RFECV — scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html. [Accessed 20-05-2024].

[29] TensorFlow — tensorflow.org. https://www.tensorflow.org/?hl=es-419. [Accessed 20-05-2024].

[30] Keras Team. Keras: Deep Learning for humans — keras.io. https://keras.io/. [Accessed 20-05-2024].

[31] PolynomialCountSketch — scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.kernel_approximation.PolynomialCountSketch.html. [Accessed 24-05-2024].

[32] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 239–247, New York, NY, USA, 2013. Association for Computing Machinery.

[33] NMF — scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html. [Accessed 24-05-2024].

[34] Non-Negative Matrix Factorization - GeeksforGeeks — geeksforgeeks.org. https://geeksforgeeks.org/non-negative-matrix-factorization/. [Accessed 02-06-2024].

[35] Supervised vs Unsupervised vs Reinforcement - AITUDE — aitude.com. https://www.aitude.com/supervised-vs-unsupervised-vs-reinforcement/#:~:text=Supervised%20Learning%20predicts%20based%20on,patterns%20and%20predict%20the%20output. [Accessed 02-06-2024].

[36] Reinforcement learning - GeeksforGeeks — geeksforgeeks.org. https://www.geeksforgeeks.org/what-is-reinforcement-learning/. [Accessed 02-06-2024].

[37] LogisticRegression — scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. [Accessed 24-05-2024].

[38] What is a neural network? - GeeksforGeeks — geeksforgeeks.org. https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/. [Accessed 24-05-2024].

[39] 7 Types of Neural Networks in Artificial Intelligence Explained — upGrad blog — upgrad.com. https://www.upgrad.com/blog/types-of-neural-networks/. [Accessed 02-06-2024].

[40] Jack V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, 49(11):1225–1231, 1996.

[41] Soumyaa Rawat. Advantages and Disadvantages of Neural Networks — Analytics Steps — analyticssteps.com. https://www.analyticssteps.com/blogs/advantages-and-disadvantages-neural-networks. [Accessed 02-06-2024].

[42] RandomForestClassifier — scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. [Accessed 24-05-2024].

[43] Random Forest: A Complete Guide for Machine Learning — builtin.com. https://builtin.com/data-science/random-forest-algorithm. [Accessed 02-06-2024].

[44] XGBoost Documentation &x2014; xgboost 2.0.3 documentation — xgboost.readthedocs.io. https://xgboost.readthedocs.io/en/stable/. [Accessed 24-05-2024].

[45] Optuna - A hyperparameter optimization framework — optuna.org. https://optuna.org/. [Accessed 20-05-2024].

[46] GridSearchCV — scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. [Accessed 02-06-2024].

[47] Linda P. Fried, Catherine M. Tangen, Jeremy Walston, Anne B. Newman, Calvin Hirsch, John Gottdiener, Teresa Seeman, Russell Tracy, Willem J. Kop, Gregory Burke, and Mary Ann McBurnie. Frailty in Older Adults: Evidence for a Phenotype. *The Journals of Gerontology: Series A*, 56(3):M146–M157, 03 2001.

[48] Felipe Diaz-Toro, Gabriela Nazar, Claudia Troncoso, Yeny Concha-Cisternas, Ana Maria Leiva-Ordoñez, Maria Adela Martinez-Sanguinetti, Solange Parra-Soto, Nicole Lasserre-Laso, Igor Cigarroa, Lorena Mardones, Jaime Vásquez-Gómez, Fanny Petermann-Rocha, Ximena Diaz-Martinez, Carlos Celis-Morales, and ELHOC Research Consortium. Frailty index as a predictor of mortality in middle-aged and older people: A prospective analysis of chilean adults. *Int. J. Environ. Res. Public Health*, 20(2):1195, January 2023.

[49] Kenneth Rockwood, Xiaowei Song, Chris MacKnight, Howard Bergman, David B Hogan, Ian McDowell, and Arnold Mitnitski. A global clinical measure of fitness and frailty in elderly people. *CMAJ*, 173(5):489–495, August 2005.

[50] Flavio Bertini, Giacomo Bergami, Danilo Montesi, Giacomo Veronese, Giulio Marchesini, and Paolo Pandolfi. Predicting frailty condition in elderly using multidimensional socioclinical databases. *Proceedings of the IEEE*, 106(4):723–737, 2018.

[51] Jordi Amblàs-Novellas, Joan Carles Martori, Núria Molist Brunet, Ramon Oller, Xavier Gómez-Batiste, and Joan Espaulella Panicot. Índice frágil-vig: diseño y evaluación de un índice de fragilidad basado en la valoración integral geriátrica. *Revista Española de Geriatría y Gerontología*, 52(3):119–127, 2017.

[52] Manuel Abbas and Régine Le Bouquin Jeannès. Acceleration-based gait analysis for frailty assessment in older adults. *Pattern Recognition Letters*, 161:45–51, 2022.

[53] Kuang-Ming Kuo, Paul C. Talley, Masafumi Kuzuya, and Chi-Hsien Huang. Development of a clinical support system for identifying social frailty. *International Journal of Medical Informatics*, 132:103979, 2019.

[54] David Gomez-Cabrero, Stefan Walter, Imad Abugessaisa, Rebeca Miñambres-Herraiz, Lucia Bernad Palomares, Lee Butcher, Jorge D Erusalimsky, Francisco Jose Garcia-Garcia, José Carnicero, Timothy C Hardman, Harald Mischak, Petra Zürbig, Matthias Hackl, Johannes Grillari, Edoardo Fiorillo, Francesco Cucca, Matteo Cesari, Isabelle Carrie, Marco Colpo, Stefania Bandinelli, Catherine Feart, Karine Peres, Jean-François Dartigues, Catherine Helmer, José Viña, Gloria Olaso, Irene García-Palmero, Jorge García Martínez, Pidder Jansen-Dürr, Tilman Grune, Daniela Weber, Giuseppe Lippi, Chiara Bonaguri, Alan J Sinclair, Jesper Tegner, Leocadio Rodriguez-Mañas, and FRAILOMIC initiative. A robust machine learning framework to identify signatures for frailty: a nested case-control study in four aging european cohorts. *GeroScience*, 43(3):1317–1329, June 2021.

[55] R.C. Ambagtsheer, N. Shafiabady, E. Dent, C. Seiboth, and J. Beilby. The application of artificial intelligence (ai) techniques to identify frailty within a residential aged care administrative data set. *International Journal of Medical Informatics*, 136:104094, 2020.

[56] Francisco M. Garcia-Moreno, Maria Bermudez-Edo, José Luis Garrido, Estefanía Rodríguez-García, José Manuel Pérez-Mármol, and María José Rodríguez-Fórtiz. A microservices e-health system for ecological frailty assessment using wearables. *Sensors*, 20(12), 2020.

[57] Andreas Philipp Hassler, Ernestina Menasalvas, Francisco José García-García, Leocadio Rodríguez-Mañas, and Andreas Holzinger. Importance of medical data preprocessing in predictive modeling and risk factor discovery for the frailty syndrome. *BMC Med. Inform. Decis. Mak.*, 19(1):33, February 2019.

[58] Manuel Abbas and Régine Le Bouquin Jeannès. Acceleration-based gait analysis for frailty assessment in older adults. *Pattern Recognit. Lett.*, 161:45–51, September 2022.

[59] Sylvia Aponte-Hao, Sabrina T Wong, Manpreet Thandi, Paul Ronksley, Kerry McBrien, Joon Lee, Mathew Grandy, Dee Mangin, Alan Katz, Alexander Singer, Donna Manca, and Tyler Williamson. Machine learning for identification of frailty in canadian primary care practices. *Int. J. Popul. Data Sci.*, 6(1):1650, September 2021.

[60] Shahidan Idris and Nasreen Badruddin. Classification of cognitive frailty in elderly people from blood samples using machine learning. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4, 2021.

[61] Muhammad Zeeshan Arshad, Dawoon Jung, Mina Park, Hyungeun Shin, Jinwook Kim, and Kyung-Ryoul Mun. Gait-based frailty assessment using image representation of imu signals and deep cnn. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1874–1879, 2021.

[62] Muhammad Zeeshan Arshad, Dawoon Jung, Mina Park, Hyungeun Shin, Jinwook Kim, and Kyung-Ryoul Mun. Gait-based frailty assessment using image representation of imu signals and deep cnn. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1874–1879, 2021.

[63] 1.13. Feature selection — scikit-learn.org. `https://scikit-learn.org/stable/modules/feature_selection.html`. [Accessed 02-06-2024].

[64] Finneas J R Catling and Anthony H Wolff. Temporal convolutional networks allow early prediction of events in critical care. *J. Am. Med. Inform. Assoc.*, 27(3):355–365, March 2020.

[65] Bryan P Bednarski, Akash Deep Singh, Wenhao Zhang, William M Jones, Arash Naeim, and Ramin Ramezani. Temporal convolutional networks and data rebalancing for clinical length of stay and mortality prediction. *Sci. Rep.*, 12(1):21247, December 2022.

[66] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. Big Data*, 8(1):53, March 2021.

[67] Google Colab — colab.research.google.com. `https://colab.research.google.com/github/kmkarakaya/ML_tutorials/blob/master/Conv1d_Predict_house_prices.ipynb#scrollTo=c26juK7ZG8j-`. [Accessed 24-05-2024].

[68] The English Longitudinal Study of Ageing (ELSA) — elsa-project.ac.uk. `https://www.elsa-project.ac.uk/`. [Accessed 24-05-2024].

[69] Sachinsoni. Different Metrices in Machine Learning for Measuring performance of Classification Algorithms — sachinsoni600517. `https://medium.com/@sachinsoni600517/different-metrices-in-machine-learning-for-measuring-performance`. [Accessed 24-05-2024].

[70] Harikrishnan N B. Confusion Matrix, Accuracy, Precision, Recall, F1 Score — medium.com. `https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd`. [Accessed 24-05-2024].

[71] Alejandro Álvarez-Bustos, Beatriz Rodríguez-Sánchez, Jose A Carnicero-Carreño, Walter Sepúlveda-Loyola, Francisco J Garcia-Garcia, and Leocadio Rodríguez-Mañas. Healthcare cost expenditures associated to frailty and sarcopenia. *BMC Geriatr.*, 22(1):747, September 2022.

[72] Iranzu Mugueta-Aguinaga and Begoña Garcia-Zapirain. Is technology present in frailty? technology a back-up tool for dealing with frailty in the elderly: A systematic review. *Aging and disease*, 8(2):176–195, 2017.

[73] L P Fried, C M Tangen, J Walston, A B Newman, C Hirsch, J Gottdiener, T Seeman, R Tracy, W J Kop, G Burke, and M A McBurnie. Frailty in older adults: Evidence for a phenotype. *J. Gerontol. A Biol. Sci. Med. Sci.*, 56(3):M146–M157, March 2001.