Università degli Studi di Napoli Federico II
Ph.D. Program in
Information Technology and Electrical Engineering
XXXVI Cycle

Thesis for the Degree of Doctor of Philosophy

Evaluating Artificial Intelligence Capabilities to Support Clinical Decisions on the Path Toward Healthcare 4.0

by
Sayna Rotbei

Advisor: Prof. Alessio Botta

Scuola Politecnica e delle Scienze di Base
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione

To all those who inspired, believed in, and supported me

Evaluating Artificial Intelligence Capabilities to Support Clinical Decisions on the Path Toward Healthcare 4.0

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy
in Information Technology and Electrical Engineering
by

Sayna Rotbei

January 2024

Approved as to style and content by

Prof. Alessio Botta, Advisor

Università degli Studi di Napoli Federico II
Ph.D. Program in Information Technology and Electrical Engineering
XXXVI cycle - Chairman: Prof. Daniele Riccio

http://itee.dieti.unina.it

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Information Technology and Electrical Engineering is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.
Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli, April 11, 2024

Sayna Rotbei

Abstract

Healthcare faces obstacles that hinder optimal care. Healthcare 4.0 provides real-time personalized healthcare services with state-of-the-art technologies. Artificial Intelligence (AI) is a critical technique for accurately diagnosing and predicting health issues. Even though AI has outperformed human diagnosis in several fields, its reliability remains unclear. Utilizing AI in this field requires understanding whether the technology is mature enough to support clinicians in their daily activities.

In order to evaluate the potential of AI, in this thesis, four distinct areas of healthcare were studied, each presenting unique challenges and opportunities. The work started with assessing the potential of AI to support doctors to have an insight into the future life quality of prostate cancer patients after prostatectomy. Then, psychiatric diseases were investigated, widespread issues of recent years, evaluating the potential of this technology to support clinicians in understanding the impact of treatment methods or lifestyle. Another important chronic disease, specifically diabetes, was then studied to understand if AI can support physicians in predicting glycemia events. This was a collaborative work with San Carlos Clinical Hospital in Madrid. The cardiology area was finally considered to evaluate the potential of Deep Learning (DL) methods for supporting clinicians to interpret Electrocardiogram (ECG) signals and relate them to cardiovascular issues. This was a joint work with Warwick University. The experiments conducted in these areas show accuracy values ranging from 63% to 95%. These results demonstrate the potential of AI in supporting clinicians toward healthcare 4.0. However, a crucial prerequisite for obtaining high accuracy is the availability of large and well-organized datasets. AI has also shown the ability to identify complex and sometimes hidden relationships among variables considered in the studies, even without a properly arranged and large dataset. It has also been able to identify the most crucial variable for gaining insights into a patient's future.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Healthcare 4.0

Sintesi in lingua italiana

L'assistenza sanitaria incontra ostacoli che rendono difficili cure ottimali. Il paradigma Healthcare 4.0 prevede servizi sanitari personalizzati real-time con l'uso di tecnologie dello stato dell'arte. L'Intelligenza Artificiale (IA) è una tecnica fondamentale per diagnosi accurate e per la previsione di problemi di salute. Sebbene l'AI abbia dimostrato vantaggi rispetto alle diagnosi umane, la sua affidabilità è poco chiara. L'utilizzo di IA in questo settore richiede di capire se la tecnologia sia matura per dare supporto agli operatori sanitari nelle loro attività. Per valutare il potenziale dell'IA, in questa tesi sono state studiate quattro distinte aree cliniche, ognuna con sfide e opportunità uniche. Il lavoro è iniziato con la valutazione del potenziale dell'IA per supportare i medici nella previsione della qualità di vita dei pazienti affetti da cancro alla prostata dopo prostatectomia. Successivamente, sono state investigate le malattie psichiatriche, un problema molto diffuso negli ultimi anni, valutando il potenziale di questa tecnologia nel supporto al personale sanitario per comprendere l'impatto dei metodi di trattamento o degli stili di vita. Un altro importante disturbo cronico, il diabete, è stato successivamente studiato per capire se l'IA possa aiutare a prevedere crisi glicemiche. Questa indigine è stata condotta in collaborazione con l'ospedale San Carlos di Madrid. Infine è stata considerata la cardiologia clinica per valutare il potenziale dei metodi di Deep Learning (DL) per aiutare i medici a interpretare elettrocardiogrammi e collegarli a problemi cardiovascolari. Questo lavoro è stato svolto in collaborazione con l'Università di Warwick. Gli esperimenti condotti in questo lavoro mostrano valori di accuratezza compresi tra il 63% e il 95%. Questi risultati dimostrano il potenziale dell'IA per sostenere il personale sanitario verso il modello Healthcare 4.0. Tuttavia, un prerequisito fondamentale per ottenere accuratezza elevata è la disponibilità di dataset estesi e ben organizzati. L'IA ha anche dimostrato la capacità di identificare relazioni complesse e a volte nascoste tra i dataset considerati in questi studi, anche senza un dataset non adeguatamente esteso e organizzato. Si è anche dimostrata capace di identificare la variabile più importante per ottenere informazioni sul futuro dei pazienti.

Parole chiave: Intelligenza Artificiale, Machine Learning, Deep Learning, Assistenza Sanitaria 4.0

Contents

Acknowledgements

I am incredibly grateful to my supervisor, Prof. Alessio Botta, who believed in me at the initial step of my Ph.D. journey and supported me constantly with his valuable advice. He made my second family here and became my role model in life. I am grateful forever for every single advice and help he gave me and his continuous support, both during my successes and in my sad moments.

I am immensely grateful to the 'Bottini' research group for their unwavering support, kindness, and expertise throughout my Ph.D. journey. Their warm welcome and kindness made my experience here in Italy truly unforgettable. They were the members of my second family. For sure by having them always around, I could tackle all challenges during my Ph.D. Thanks for the beautiful memories, the contagious laughter, and the cherished time we spent together.

I am truly grateful for the expert guidance and warm welcome extended to me by Prof. Giuseppe Fico and the Life Support Technology group of the Polytechnic University of Madrid, Spain, Where I spent my period abroad. Their invaluable contributions have been instrumental in the success of my research. The time I spent there was truly unforgettable and will always hold a special place in my heart.

I would like to express my gratitude to Prof. Daniele Riccio for his support and coordination.

I owe my beloved family immense gratitude for showing me the path to success. Your unwavering support and teachings have been my greatest inspiration in life. Your belief in me has fueled me never to give up and strive for the best. Dear Mom, I want to take a moment to express my gratitude towards you for being my pillar of strength and guidance. Your unwavering support and wise counsel have been a source of inspiration and motivation for me. Thank you for being a shining example of love and care. Dear Dad, thanks to your courage, I have learned from you to approach difficult circumstances more positively and constructively. My dearest friend and brother Farzad, I feel incredibly lucky to have you in my life as an older brother; thanks for being my safe person. Your love, kindness, and endless support have downgraded all my difficulties during the last three years when I have been far from home. I am grateful for how you inspire me to be my best self. This accomplishment belongs to you all just as

much as me. Thank you for always being there for me.

I would like to express my sincere gratitude to every member of Arclab, and to all my dear friends. Your unwavering support, encouragement, and kindness have played an instrumental role in shaping my personal and professional growth. Your presence in my life has been a source of inspiration and motivation, and I am truly grateful for every moment shared with you. Thank you for being a part of my journey and for filling my life with warmth and love.

List of Acronyms

The following acronyms are used throughout the thesis.

NHS	National Health Service
ML	Machine Learning
AI	Artificial Intelligence
RF	Random Forest
SVM	Support Vector Machine
kNN	k-Nearest Neighbors
LRM	Logestic Regression Model
LR	Linear Regression
MLP	Multi-Layer Perceptron
DT	Decision Tree
NB	Naive Bayes
ANN	Artificial Neural Network
DL	Deep Learning
DNN	Deep Neural Network
LSTM	Long Short-Term Memory
ROB	Risk of Bias

XGBoost	Extreme Gradient Boosting
GBoost	Gradient Boosting
SGD	Stochastic Gradient Descent
AdaBoost	Adaptive Boosting
CDSSs	Clinical Decision Support Systems
EHRs	Electronic Health Records
AI-DSS	AI-Driven Decision Support Systems
PCa	Prostate Cancer
HRQOL	Health Related Quality of Life
AUC	Area Under the Curves
LRM	Logistic Regression Model
LR	Logistic Regression
RP	Radical Prostatectomy
EPIC	Expanded Prostate Index Composite
IQR	Interquartile range
IPSS	International Prostate Symptom Score
IIEF	International Index of Erectile Function
tDCS	Transcranial Direct Current Stimulation
HDRS	Hamilton Depression Rating Scale
DM	Diabetes Mellitus
T1D	Type 1 Diabetes
T2D	Type 2 Diabetes
CGM	Continuous Glucose Monitoring
KET	Key Enabling Technologies
HR	Heart Rate

OCD	Obsessive Compulsive Disorder
AD	Adjustment Disorder
Y-BOCS	Yale-Brown Obsessive-Compulsive Scale
BABS	Brown Assessment of Belief Scale
BDI-II	Beck Depression Inventory-II
STAI-Y	State-Trait Anxiety Inventory-Y
tDCS	Transcranial Direct Current Stimulation
HDRS	Hamilton Depression Rating Scale
T1DM	Type 1 Diabetes Mellitus
T2DM	Type 2 Diabetes Mellitus
ML	Machine Learning
RF	Random Forest
AI	Artificial Intelligence
GDP	Gross Domestic Product
OECD	Organisation for Economic Co-operation and Developmen
EU	European Union
DM	Diabetes Mellitus
T2D	Type 2 Diabetes
WHO	World Health Organization
KNN	KNearest Neighbours
NB	Naive Bayes
SVM	Support Vector Machine
DT	Decision Tree
LR	Logistic Regression
AUC	Area Under the Curve

SHAP	SHapley Additive exPlanations
CDF	Cumulative Distribution Function
PDP	Partial Dependence Plots
ECG	Electrocardiogram
NN	Neural Network
RBFN	Radial Basis Function Networks
NLPCA	Non-Linear Principal Component Analysis
HMM	Hidden Markov Model
DL	Deep Learning
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
EKF	Extended Kalman Filter
AR	AutoRegressive
FP	False Positive
GMM	Gaussian Mixture Model
KFs	Adaptive Kalman Filters
FPs	fiducial points
ML	Machine Learning
LR	Linear Regression
CNN	Convolutional Neural Network
AI	Artificial Intelligence
RNNs	Recurrent Neural Networks
ConvLSTM	Convolutional Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
MultiHMM	Multi Hidden Markov Model

PCGS Partially Collapsed Gibbs Sampler

RMSE Root Mean Square Error

Conv-BiLSTM Convolutional Bidirectional Long Short Term Memory

FP Fiducial Point

STFT Short-Time Fourier Transform

List of Figures

List of Tables

Chapter 1

Introduction

Managing patient care requires knowledge, skills, and responsibility for a vast amount of medical information. In addition to the complexity and uncertainty of the healthcare system, healthcare managers and clinicians face many challenges, like diagnosing, deciding the treatment method in a limited time, providing consultation, predicting the side effects of different treatment methods, and preparing for them. Healthcare systems globally assumed goals, such as improving population health, enhancing patient and caregiver experiences, and reducing care provision costs [17]. It is also expected to provide effective, high-quality care and implement real-world data insights directly in patient care to achieve transformation at scale. Therefore, there is a growing need for innovation and transformation in healthcare delivery models [17]. The pandemic has further highlighted the shortages in the healthcare workforce and the inequities in access to care, which were previously identified by The King's Fund and the World Health Organization [52]. It is estimated that there will be almost 250,000 vacancies in National Health Service (NHS) trusts by 2030 due to a shortage of staff [52]. By 2030, the global population will face a shortage of 18 million healthcare professionals, including five million doctors, and developing countries will be affected the most [118].

The implementation of technology and AI in healthcare has the potential to tackle some of the supply-and-demand challenges. Recently, healthcare providers like hospitals and clinics are increasingly using Electronic Health Records (EHRs) for medical records, which provide humongous amounts of data [146]. This creates a high potential for AI-based technologies to provide better services and reduce medical errors [111]. AI-based applications can identify and highlight correlations between large amounts of data that would otherwise have escaped the attention of clinicians and researchers. They can use different sets of demographic and clinical data and information related to the patients in order to provide diagnoses, predictions, or treatment recommendations as output [28]. There is an expectation of a 29.3% growth rate in AI applications in healthcare by 2025, as well as a 40% increase in global revenue [146]. In the future, the healthcare system is expected to shift towards preventive medicine and AI outpatient clinics due to the abundance of patient data. Moreover, AI applications are capable of providing more accurate and reliable clinical decisions, and as a result, they may become an integral part of the healthcare system [146].

The main target of Healthcare 4.0 is to assist the specialists whenever they require it. The techniques and methods related to Healthcare 4.0 can provide the information needed to help specialists make better decisions. The objective of the initiative is to aid patients and their family members in efficiently managing their healthcare requirements [111]. Figure 1.1 illustrates the primary tools employed in Healthcare 4.0 and the accomplishments it has achieved.

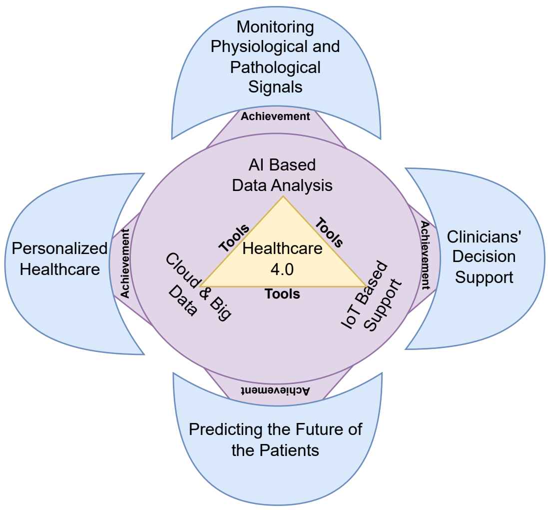

Figure 1.1. Tools and achievements of Healthcare 4.0

Although AI systems in the field of healthcare 4.0 like Clinical Decision Support Systems (CDSSs) are enhancing the quality, safety, reducing the adverse effects of medications, and efficiency of the hospital and patient care, there are still some challenges that needs to be address [128]. The first question about them is their trustworthiness: are AI-based methods trustable, and what does it take to advance towards trustable systems and implement and realize trustworthy AI in the clinic? In order to answer this question, this thesis evaluated different AI methods, including ML and DL methods, in different areas, not only to assess them but also to suggest trustable methodologies to use on the path towards healthcare 4.0. Moreover, it aims to explore the potential of AI techniques to empower physicians and clinicians in delivering improved healthcare services to patients in the context of healthcare 4.0. To achieve this objective, we collaborated with various clinical departments and universities, namely the urological department of the University Hospital of Naples, Federico II, and Salerno; the psychiatric department of the University Hospital of Naples, Federico II; an emergency section of the San Carlos Clinical Hospital; the University of Warwick; and the Universidad Politécnica de Madrid.

As the first contribution, urological patients were considered. These days, AI becomes increasingly common methods to be applied to diagnose, treat, and predict urological conditions [146]. In this field, AI techniques are applied to improve cystoscopic diagnosis and prognosis and survival prediction of bladder cancer patients [81]. Moreover, ML -based prediction algorithms and models are being developed in order to identify genes that have the potential to predict the recurrence or future progression of Prostate Cancer (PCa) [146]. This disease is a significant public health issue owing to its high incidence rate in men, which affects human life quality. Prostatectomy surgery is one of the common and effective treatment options. Therefore, the primary objective of this contribution was to evaluate the potential of ML algorithms for predicting patient quality of life after 12 months since prostatectomy. We focused specifically on the sexual function as self-reported by patients. Baseline and one-year patient demographic and clinical data formed the basis of our case study. In this study, Expanded Prostate Index Composite (EPIC)-26 questionnaire was utilized to allow individuals to assess their sexual function. Based on the results obtained, the use of AI methods enabled the prediction of patients' self-assessment of their sexual life and the identification of important predictors. By employing the results obtained, clinicians are able to provide better consultations to patients and their families about their future. Moreover, clinicians can decide wisely about the treatment method for patients.

As the next clinical domain, we collaborated with the University Hospital of Naples, Federico II psychiatric department. According to the Global Burden of Disease Study in 1990 and 2010, mental and substance-use disorders are the fourth largest disease burden measured in disability-adjusted life years (DALYs) [99]. These disorders are considered the foremost cause of lived-with disability worldwide [99]. In recent years, to diagnose and predict psychiatric disorders, AI-based technologies are widely employed [82, 48]. Different AI approaches hold great potential in improving the efficiency and accuracy of diagnosis, treatment recommendations, predictive interventions, and allocation of scarce resources in psychiatry [31]. Considering all the above points, we worked on two different branches of this field to evaluate AI-based methods on the path toward

Healthcare 4.0. The first one was about knowing the effect of large-scale dramatic events such as pandemics and war on psychiatric symptoms. In this case, we create a methodology based on supervised ML to identify predictors of the severity of psychiatric symptoms during the COVID-19 lockdown as well as to predict the degree of severity. Then, we applied this methodology to a dataset of real cases to evaluate its reliability. The second branch of this collaboration was about using AI for investigating the impact of Transcranial Direct Current Stimulation (tDCS) treatment method on depressed patients. We used Hamilton Depression Rating Scale (HDRS) tool to measure the level of patients' depression. The purpose of this collaborative work was to determine if ML methods are able to predict HDRS level ten/twenty sessions following the start of treatment. The result of this study also provides interesting insight into the factors that contribute the most to the predictions.

We then focused on diabetes, in a joint project with Universidad Politécnica de Madrid and the University of Warwick. From 1980 to 2017, the number of people with diabetes rose from 108 million to 425 million globally, and it is expected to reach 629 million in 2045 [65]. There are numerous complications and significant morbidity and mortality associated with diabetes, which makes early detection and prevention, as well as treatment and management of diabetes, essential [47]. Self-management is widely acknowledged as essential for reducing the risks of chronic complications in people with diabetes [96]. Technological advancements have enabled wearables, smartphones, and other gadgets to monitor and track symptoms, disease status, and diabetic management activities [47]. Participating in most types of physical activity is beneficial for managing diabetes; studies have demonstrated the importance of physical activity in preventing and treating diabetes [75]. We, therefore, conducted a systematic literature review to investigate how data-driven AI-based methods can improve diabetes management using tracking data from Key Enabling Technologies (KET). This review aimed to provide an understanding of how technology-based monitoring of physical activity can assist with diabetes management. After understanding the role of AI-based methods in the diabetic field, we focused on building a methodology for early detection of glycemia events in the patients of the emergency section. This work investigated data from the emergency department of the San Carlos Clinical Hospital in the Madrid region of Spain to predict hyperglycemic and hypoglycemic events that may occur within the hospital setting, as well as to identify the role of different characteristics of patients on such events. In this scenario, healthcare professionals are empowered to offer improved treatment methods that are highly effective in mitigating glycemic events and their associated risks and complications.

As the last healthcare domain, we focused on cardiovascular diseases. As ECG signal provides critical cardiac information, researchers analyze it to detect and prevent life-threatening heart problems [171]. Therefore, we examined AI-based methods to analyze the ECG signals. This signal provides information regarding the electrical activity of the human heart, as well as ECG signal morphologies that indicate the type of arrhythmia associated with those conditions [171]. Accurate segmentation and detection of fiducial points are crucial to analyzing ECG signals and detecting certain cardiac conditions [69]. Currently, clinicians rely on manual analysis of ECG signals to detect heart-related conditions. However, with advancements in digital technology and wearables for monitoring ECGs and other physiological parameters, the design and development of automated ECG segmentation tools have gained significant attention [70]. Thus, we focused on diagnosing the health status of the patient through the analysis and classification of ECG signals. This work, in collaboration with the University of Warwick, aimed to explore different DL models and methodologies for precisely segmenting ECG signals. The work was conducted on the PhysioNet's QT public dataset in order to find the most effective technique for automating ECG signals analysis. Different methods were tested for classifying ECG signals into heartbeat parameters, obtaining promising results. This shows that AI-based methods have the potential to support clinicians in this important field.

This thesis explores how clinicians and physicians can enhance the quality of their clinical and healthcare services, achieve their goals, and overcome current and future challenges using AI-based methods in different clinical and healthcare areas. It will not only assist them in making well-informed decisions but also help them to be efficient in the way they spend their time.

1.1 Used Material and Methods

In order to explore the plausibility of leveraging AI-powered methods to aid clinicians in their decision-making processes, we conducted a thorough examination of a variety of such methods across different use cases. Our goal was to gain a comprehensive understanding of the potential benefits and limitations of utilizing AI technology in clinical settings. Figure 1.2 depicts the commonly applied methods and techniques for all domains.

Before launching AI algorithms, data has been pre-processed in order to handle imperfections (which may include data with errors, noise, or missing values) [92]. So, in all used datasets, all data types have been replaced by numerical values, e.g., date of birth, and missing values are converted to age or minus one, respectively. After managing all imperfections in the dataset and ensuring their unifying, datasets were normalized to be ready for applying different AI methods.

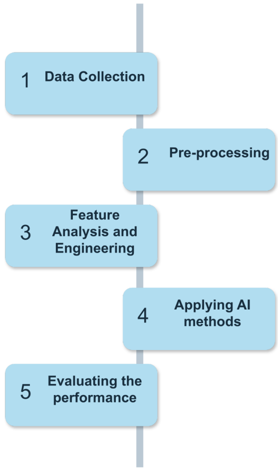

Figure 1.2. Applied workflow for all domains

1.1.1 Feature Analysis and Engineering

Prior to starting the classification process, it is imperative to understand the role of different predictors (also referred to as features) and why they are important for the classification process [34]. Consequently, we will be able to ensure that the information we collect is relevant, meaningful, and formatted to allow us to accomplish the goal at hand as effectively and efficiently as possible. Moreover, we will be able to achieve more accurate and more credible results. In addition, it may be possible to reduce the size of a dataset without compromising its quality by identifying the importance of different features contained in the dataset [34]. Further, an in-depth understanding of the features will also facilitate the choice of the most appropriate ML algorithm and the development of effective features. It also means that the findings relating the target variables to the features are another important outcome of feature analysis [136]. Based on the importance of each feature, the order in which they are considered is determined.

To conduct this study, we employed a variety of analytical methods that had been tailored to address the aim of each particular healthcare domain. We utilized the SHapley Additive exPlanations (SHAP) method as one of our instance-based explanation methods [26]. This helped us to provide explanations about the ML methods that we used and the role of variables in our analysis [26]. In order to thoroughly analyze the variables and their contribution towards achieving the target goal, it is advisable to plot the correlation among all predictors of the dataset. Although, for space reasons, I did not visualize the correlation matrix for all the datasets, I checked the correlation among predictors in all datasets. Moreover, to analyze the features further, the decision tree method has been implemented to effectively demonstrate the contribution made by each predictor in predicting the target value [72]. Decision tree algorithms like classification and regression trees provide importance scores based on split point criteria. This approach works for tree ensembles, such as Random Forest and Stochastic Gradient Boosting. To enhance the accuracy of the classification model, we applied feature engineering to extract and construct features that are specifically designed for analyzing ECG signals. All methods used were highly effective in providing valuable insights about the dataset and utilized AI methods to make informed decisions.

1.1.2 Machine Learning Models Used

The classification experiments were conducted employing a variety of ML models. The performance analysis presented in the thesis pertains to the outcomes of the diverse models described below. The study aims to analyze the predictive capabilities of these models and assess their performance. These findings are expected to contribute to provide valuable insights into the practical application of ML algorithms.

Random Forest (RF) is useful for classification and regression tasks. The method combines unreliable multiclass classifiers to reduce variance and produce more reliable results. RF can detect outliers by calculating pair closeness and reducing variance without affecting prediction accuracy [151]. The performance of each classification was assessed by comparing the results of each tree [90]. The following method used was the Support Vector Machine (SVM) algorithm. It groups data to find the best boundary between classes and creates complex models by transforming input data into a higher dimensional space [90, 116]. Concerning the feature space, SVMs utilize the segment of training data that is closest to the finest decision boundary [151]. Another method is k-Nearest Neighbors (kNN) method, which has been described as one of the top 10 data mining techniques [115]. The algorithm uses the classification of adjacent samples to determine the classification of an unknown sample [115]. It calculates the distance between an unknown sample and all the samples in a training set to find the closest sample. Decision Tree (DT) is a tree-like structure where a separating sequence describes the path from root to leaf node until a Boolean outcome is reached [32]. It depicts knowledge relationships hierarchically with nodes and connections, classifying and representing purposes [32]. Linear Regression (LR) is another used ML method. To determine the best regression coefficient, the training classifier employs an optimization algorithm [182]. During current work, the Multi-Layer Perceptron (MLP) method is used. It is a feedforward neural network that has one or more layers between its input layer and its output layer [4]. This algorithm is commonly utilized for tasks such as pattern recognition, classification, approximation, and prediction [4]. Stochastic Gradient Descent (SGD) is used in most of the domains in the current research. Based on a stochastic approximation of gradient descent optimization, it iteratively optimizes a differentiable objective function [119]. The SGD algorithm can handle large amounts of data by breaking up data into smaller portions. The loss function is optimized using gradient descent and assumes independent and uniform samples [15]. Adaptive Boosting (AdaBoost) is another used ML methods in this work. Through the use of sequential training, AdaBoost builds a composite classifier that emphasizes specific patterns more and more [143]. In regression problems, the algorithm fits the dataset through a regressor and then fits additional instances using different weights. AdaBoost is commonly used in classification and regression [157]. The Gradient Boosting (GBoost) is one of the usedML methods, which is an ensemble ML technique used for regression and classification [16]. It minimizes a loss function by moving in the opposite direction of the gradient. The loss function measures the accuracy of the model and varies by problem type [16]. Boosting involves building sub-trees sequentially from an original tree, reducing errors with each subsequent tree [41]. A dummy classifier is also used in one of the research domains as a baseline for comparing the performance of ML algorithms. It generates output labels based on the statistical distribution of the dataset [108].

For evaluating the performance of the models used, k-fold cross-validation was applied to obtain more stable and generalizable results and avoid overfitting. This assessment method divides the datasets into k equal subsets of individuals called folds. Training is conducted on K-1 of these folds, followed by validation on the remaining fold. The training and validation sets were changed once in each fold, and this process was repeated K times. Lastly, we tested each of the k subsets once as a test set to determine how well the model performed [43].

Two main types of classification tasks, Binary classification, and Multi-class classification, are employed based on kinds of target values and expectations. Binary classification categorizes a given dataset into two distinct classes and predicts each sample belongs to which class [83]. On the other hand, multi-class classification categorizes the dataset into multiple classes based on a classification rule [83]. In this work, we used both methods mentioned. Moreover, to assess the behavior of all algorithms when they were faced with the dataset, I checked their confusion matrix, although, for space reasons, I do not provide all of them in this thesis.

1.1.3 Performance Metrics

To determine the effectiveness of the classifiers, it is essential to evaluate their performance. In this thesis, four metrics are considered for evaluation: precision, recall, F1-score, and accuracy.

The number of false-positive (FP), false-negative (FN), true-positive (TP), and true-negative (TN) are used for obtaining accuracy, which is the ratio of correct predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The number of FP, FN, and TP are used for calculating the precision and recall which are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

As a result of using the F1-score, datasets with an imbalanced distribution are more likely to be analyzed better [165]. It can be calculated by using the ratio of recall and precision or TP, FP, and FN as follows:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

1.2 Thesis Outline

After the introduction chapter, the thesis proceeds as follows: in chapter 2, the dataset related to the urology patients is considered to predict their life quality after prostate cancer surgery in terms of self-assessment of patients. In the next step, the databases related to psychiatric patients are covered in chapter 3 to evaluate the potential of AI-based methods in supporting clinicians in their decision-making processes. Chapter 4 is focused on the diabetic issue in terms of reviewing current AI-based technology and providing a method for predicting glycemic events. As another work domain ECG signals are classified into heartbeat parameters in chapter 5, and finally, the conclusion of the whole work is provided in chapter 6.

Chapter 2

Urology

2.1 Introduction

The most common malignant disease in men is PCa, which currently causes the third highest number of cancer deaths in men in Europe [63]. PCa accounts for 1,276,106 new cases yearly and caused 358,989 deaths (3.8% of all deaths caused by cancer in men) in 2018.

Both the patient and their partner are affected by thoughts of death and dying when faced with a life-threatening diagnosis [141]. Urological surgery is a common treatment used to treat most prostate cancers effectively. Nevertheless, studies indicate that shortly following the operation, patients report that erectile dysfunction is a serious problem [141]. It may be due to the fact that erectile dysfunction following radical prostatectomy impacts not only the patient's sexual life but also their sense of self and relationships with others.

Predicting the conditions of patients after surgery by using AI-based methods is beneficial to know the Health Related Quality of Life (HRQOL) of patients in the future. It not only helps to predict the quality of life of patients to counsel them better, but it is also crucial to predict potential side effects of urological surgery according to patient profile and situation before surgery.

It is fundamental to have a straightforward, easy-to-use tool for assessing HRQOL of the patients [162], such as clinical AI-Driven Decision Support Systems (AI-DSS). One of the proven methods to assess the condition of patients is the EPIC-26. EPIC-26 is a simplified version of the EPIC-50 questionnaire which contains a set of 50 questions. Due to the need to facilitate the measurement of HRQOL for use in more comprehensive research and clinical approaches, Szymanski et al. narrowed it to 26 questions the EPIC-50 questionnaire to develop the EPIC-26 [152]. Like the full version of EPIC, this version contains items belonging to five domains: urinary incontinence, irritative/obstructive urinary symptoms, bowel symptoms, sexual symptoms, hormonal symptoms, and multi-item scores quantifying symptoms between the above domains.

The increasing accessibility of data such as the EPIC questionnaire in health-care has made systems based on AI a critical approach for analyzing patient data and extracting useful information. ML, in fact, is one of the most popular methods of AI for identifying complex patterns, trends, and correlations between variables. It can easily handle multidimensional data sets and identify the most effective predictors [44]. In this regard, Gravina et al. first evaluated the ability of ML methods to identify prostate cancer [63]. The authors discovered that ML classifiers may be helpful to physicians when making diagnostic decisions. ML was also used in the study of bicalutamide, a standard treatment for prostate cancer, and its effect was predicted from data belonging to three phase III clinical trials [19]. Results prove that tools based on ML can be used to predict cancer treatment outcomes and improve the efficiency of clinical-stage drug development. Prostate cancer risk can also be predicted by using biochemical parameters with dense neural networks model [126]. The aim of another study was to predict the occurrence of erectile dysfunction at 1 and 2 years after diagnosis using patient demographics, clinical data, and patient-reported outcomes collected at diagnosis. To train and validate models, the study used a subset of 964 cases of localized prostate cancer from 69 Dutch hospitals of the Netherlands Comprehensive Cancer Organization. The logistic regression algorithms were combined with Recursive Feature Elimination to create the models for achieving expected goals [74]. An independent study sought to identify the main variables predictive of health-related quality of life in men with prostate cancer at one year. Based on 433 measurements performed in each patient, they analyzed data from 2670 patients. The aim of this study was to improve the interpretation of the data by identifying the baseline variables that can predict HRQOL at one year. From this study, it appears that the use of ML techniques is essential to identify potential predictive variables for HRQOL one year after the completion of the study [149]. Pan et al, in their research, adopted the AUC and ML methods to identify latent dosimetric parameters associated with patient-reported outcomes after stereotactic prostate irradiation [124].

In another work, a model to predict sexual function after Radical Prostatectomy (RP) was developed, which was reliable enough to determine the recovery of sexual function 12 and 24 months after RP [6]. Providing that a dynamic, multivariate model (with multiclass outcomes) can be used for decision-making in the survival phase, reducing the possibility of regretting a decision. In this study, three outcomes related to sexual function were assessed (1) the EPIC-26 sexual domain score (range 0-100); (2) the EPIC-26 sexual domain score dichotomized at ≥ 73 for 'good' function; and (3) a dichotomized variable for the quality of erection 12 and 24 months after RP [6]. A gradient-boosting decision tree was used for the prediction model. To evaluate the performance of the model, root means squared error (RMSE) and mean absolute error (MAE) were used, and the EPIC-26 score was used as continuous variables. The area under the receiver operating characteristic curve was applied for the dichotomized EPIC-26 sexual domain score and erection quality [6]. Another study examined EPIC-26 domain scores to identify the most significant features using Cluster-based Random Forest [148].

Current research aims to find a method to predict the post-operative condition of patients based on their pre-operative condition in order to evaluate the capability of AI-based method in the path toward healthcare 4.0.

2.2 Materials and Methodology

The main goal of this study is to predict patient satisfaction with sexual function after 12 months since a prostatectomy surgery with statistical and ML algorithms to use in AI-based systems such as CDSSs.

In our study, HRQOL is represented by the answer to question 12 (Q12) of the EPIC-26 questionnaire administrated before and after surgery to patients. EPIC-26 question number 12 is one of the most important components of the evaluation process, as it specifically asks: *EPIC-26-Q12: Overall, how big of a problem have your sexual functions been for you during the last 4 weeks?*

- *1: No problem.*
- *2: Very small problem.*
- *3: Small problem.*
- *4: Moderate problem.*
- *5: Big problem.*

In spite of the fact that EPIC-26-Q12 does not provide information about the sexual function of a patient from a medical standpoint, it is nevertheless capable of providing information about patient satisfaction regarding their sexual function and how patients perceive their sexual function. Therefore, it is fundamental to predict the outcome of question 12 based on the patients' profiles before surgery.

To predict Q12, we adopted data that comprises demographic and clinical data as well as responses to the EPIC-26 questionnaire before surgery. In the first step of our study, statistical methods are applied with the Pearson test [125] in order to understand the correlation between the profile of the patient and the possible outcome of question number 12. Then, in the following steps, different ML algorithms were applied to the database containing the profile of patients and their answers to the EPIC-26 questionnaire before surgery. As the first step of ML method, a multi-class classification is applied to predict the specific answers among the five possible options of EPIC-26-Q12 after 12 months. Then, binary class classification is used to predict for prediction. In this regard, the database was divided into two categories, namely those who had no problems 12 months after surgery and those who suffered from some level of problems 12 months after surgery.

2.2.1 Participants

This study included a set of 409 patients of the *Azienda Ospedaliera Universitaria* (AOU) Federico II, the Hospital of the University of Naples Federico II, Italy. The inclusion criteria were as follows: first, all patients must have undergone a surgical procedure in the field of urology; in particular, all patients were treated for prostate cancer; second, surgery must have been performed at least a year before. Patients whose surgery was less than one year ago were excluded *a priori* from the study. All examined patients underwent robot-assisted radical prostatectomy in the department of AOU Federico II from 2016 to 2021. The study was conducted according to the Declaration of Helsinki and approved by the Ethics Committee, University of Naples Federico II. Approval code 118/20.

Regarding the participants, the overall median age was 65 years (Interquartile range (IQR): 61-70), the median preoperative IPSS, and International Index of Erectile Function (IIEF) were 18 (IQR: 13-22) and 8 (IQR: 4-14), respectively. Of all, 225 (55.0%) received a nerve-spearing approach during the radical prostatectomy. Finally, after surgery, 174 (42.5%) patients did not get any medication for erectile function recovery, whereas 188 (46.0%), 29 (7.1%), and 18 (4.4%) were treated with oral therapy, oral therapy plus intracavernous injection, and only intracavernous injection, respectively. Table. 2.1 shows the main characteristics of the participants.

2.2.2 Clinical Measures and Data Collection

Among 524 subjects, 409 PCa patients who had undergone radical prostatectomy were included in the study. we removed 115 cases from the study because they did not provide an answer to question number 12 of EPIC-26 one year after surgery. The analysis could not be conducted on them. For all the patients included in the study, we performed a retrospective data collection from the patients' personal electronic medical records using a defined source hierarchy. we had 57 characteristics for each patient. Our dataset is also currently available in the urology department of AOU Federico II in Naples. The dataset contains data related to the EPIC-26 questionnaire and includes demographic data such as date of birth, information about the patient tumor, treatments, and surgeries. The EPIC-26 questionnaire has been completed at baseline (one day before surgery), and one year after surgery. All patients enrolled in this study were contacted by phone 12 months after surgery for a follow-up interview by the data manager of our center.

Table 2.1. Main characteristic of the participants

		Overall
Age	**Mean (SD)**	64.9 (0.32)
	Median	65
	IQR	61-70
IIEF	**Mean (SD)**	16.7 (0.352)
	Median	18
	IQR	13-22
IPSS	**Mean (SD)**	9.5 (0.412)
	Median	8
	IQR	4-14
Nerve sparing, n (%)	**No**	114 (27.9)
	Yes	225 (55.0)
Sexual rehabilitation therapy n (%)	**No**	174 (42.5)
	Oral therapy	188 (46)
	Oral therapy plus intracavernous injection	29 (7.1)
	intracavernous injection	18 (4.4)

2.2.3 Statistical Analysis

In this study, we analyzed data using statistical analyses traditionally used in the medical field as well as machine learning. As for the statistical analyses, two sets of them were performed, as follows. First, the Pearson test was used to assess potential correlations between the EPIC-26 question 12 and the profile of the patients (e.g., age, IIEF, IPSS, and time in months between diagnosis and surgery). Second, univariate LR was performed to predict if the response to the EPIC-26 question 12 was ≥ 3. This choice is motivated by the fact that a clinically significant worsening quality of life is associated with a value of < 3, according to the options provided in question number 12. In all statistical analyses, the R software [1] was used. All tests were two-sided with a level of significance set at $p < 0.05$. Descriptive statistics were presented as mean with standard deviation and median with IQR for continuous variables or counts and percentages for categorical variables.

2.2.4 Data Pre-processing for Machine Learning

Before launching ML algorithms, data has been pre-processed. It should be highlighted that there are no other data in the dataset related to EPIC-26-Q12

Table 2.2. Distribution of classes in case of multiclass classifications

Class_One	Class_Two	Class_Three	Class_Four	Class_Five
287	35	48	35	4

that can bias ML methods in predicting EPIC-26-Q12.

Two main types of analyses were then applied: multiclass and binary classification [62]. For the first type of analysis, only the *null* values of the answers to question 12 were removed from the dataset. The remaining values (from 1 to 5) are considered as separate classes. For the binary classification, the possible five answers of question 12 were assigned to two main groups: those who chose option number 1 were classified as group zero, and the remaining options were classified as group one. This classification separated patients who had no problems from those who had some problems. Table. 2.2 and Table. 2.3 show the distribution of classes in both cases.

Table 2.3. Distribution of classes in case of binary classification

Class_Zero	Class_One
287	122

It is important to emphasize that no data from month 12 was used to predict patient conditions at this time. This has been done in order to evaluate the capability of ML methods for predicting the future of patients only based on information available before surgery.

Classifiers were evaluated according to four metrics: precision, recall, F1-score, and accuracy [62].

2.3 Prediction Results

2.3.1 Results from Statistical Methods

In this section, we present the results of the logistic regression analysis. As a statistical method, logistic regression is used in this study for detecting the correlation between variables of the dataset and EPIC-26-Q12. Table. 2.4 shows that a statistically significant negative correlation between EPIC-26 question 12 and IIEF ($r = -0.12, p = 0.02$). A positive statistically significant correlation was recorded between EPIC-26 question 12 and IPSS ($r = 0.20$, $p \leq 0.001$), and time in months from diagnosis to surgery ($r = +0.28$, $p < 0.001$). Conversely, no statistically significant result was recorded between EPIC-26 question 12 and age ($r = 0.02$, $p = 0.4$).

Table 2.4. Correlations based on logistic regression, IPSS is significant one

Variables		R coefficient	p-value
Epic_26, question 12	**IIEF**	-0.12	0.02
	IPSS*	+0.20	0.001
	Age	0.02	0.4

As reported in Table. 2.5, according to the results of the univariate Logistic Regression Model (LRM) predicting the response to the EPIC-26 question 12 ≥ 3, IPSS (Odds Ratio [OR]: $1.04, 95\%$ confidence interval [CI]:$1.00 - 1.07, p = 0.01$) IIEF (OR: $0.96, 95\%$ CI: $0.92 - 0.99, p = 0.04$), IPSS (OR $1.04, 95\%$ CI: $1.00 - 1.07, p = 0.01$) were independent predictive factors. Age, sexual rehabilitation therapy (no vs. only oral vs. oral plus injection vs. only injection), and nerve-sparing (yes vs. no) were not independent predictor factors (all $p > 0.1$).

Table 2.5. Result of the logistic regression model, IPSS is significant one

	OR	**2.5%**	**97.5%**	**p-value**
Age	1.0128	0.9763	1.0517	0.5
IIEF	0.9635	0.9294	0.9992	0.04
IPSS*	1.0405	1.0073	1.0749	0.01
Sexual therapy reabilitation				
No	Ref			
Oral therapy	0.5494	0.3270	0.9128	0.218
Oral therapy plus intracaveronous injections	0.8854	0.3317	2.1206	0.7942
Intracaveronous injections	0.5565	0.1248	1.7823	0.3712
Nerve sparing				
No	Ref			
Yes	0.9149	0.5317	1.6013	0.7509

2.3.2 Results from Machine Learning Methods

The following section presents the results obtained with ML to predict the value of EPIC-26-Q12 at month 12.

Multiclass Classifier

Initially, multiclass classification was applied. We divided the answer to EPIC-26-Q12 in five classes as reported in Sec. 2.2.4. Figure. 2.1 illustrates the importance of the features. This figure highlights the most important predictors. The IPSS appears to be the most valuable predictor. It is a well-established method for assessing the degree of Lower Urinary Tract Symptoms (LUTS) associated with Benign Prostatic Obstruction (BPO). The IPSS can be used to determine the International Prostate Symptom Score [91]. During the survey of IPSS, eight questions are asked about both voiding and filling symptoms, in addition to how these symptoms affect a patient's quality of life [91]. The most recent Prostate-Specific Antigen (PSA) level (measured in ng/mL) in the blood and the age of the patient also have significant importance. This means that self-assessed sexual function is highly related to these factors, which constitute important predictors of such function after the surgery.

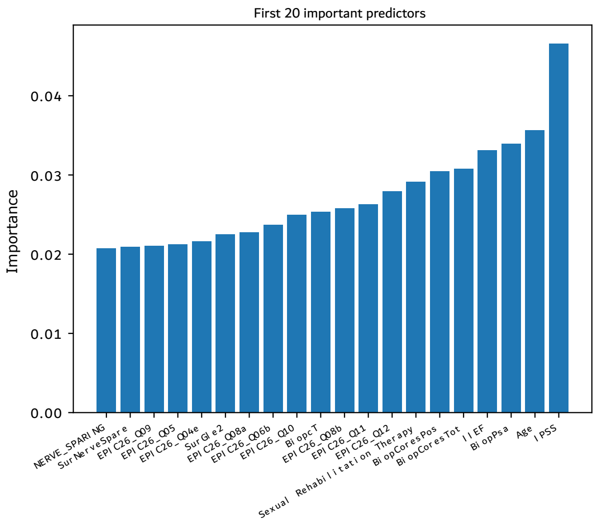

Figure 2.1. Importance of the different features in case of the multiclass classifier

Table. 2.6 depicts the performance metrics of the prediction models, including RF and SVM. We used five-fold cross-validation for this analysis. The classifiers made accurate predictions up to 71% while their F1-scores up to 61%. Recall values of up to about 71% and precision of up to about 58% are observed. Considering that we asked the algorithm to predict the EPIC-26-Q12 answer from five classes, the results are interesting in the sense that they show that such prediction is possible and accurate in some cases, even on a small-size dataset as the one considered in this work. Further investigation is necessary using a larger dataset to confirm and possibly improve this result.

Table 2.6. Result of the prediction in case of the multiclass classifier

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.58	0.71	0.61	0.71
SVM	0.49	0.70	0.58	0.70

Binary Classifier

In the subsequent step, a binary classification method was utilized. The answers to EPIC-26-Q12 have been divided into two groups, one including all patients who had no problems 12 months after surgery and the other including all patients who had any kind of problem 12 months after surgery. A comparison of patient distributions between the two classes is presented in Table. 2.3.

Figure. 2.2 shows the feature importance, i.e. how important each predictor is with respect to the value to be predicted. This figure shows that the Pathological Tumor Stage (SurT) had the greatest influence on the outcome, followed by the IPSS. The figure shows that a large set of features, besides the two cited before, has similar levels of importance for EPIC-26-Q12 in the binary case.

Table 2.7. Results of prediction in case of the binary classifier

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.72	0.74	0.71	0.74
SVM	0.73	0.73	0.66	0.73

A summary of the classification results is presented in Table. 2.7. Both SVMs and RFs had approximately similar accuracy levels. SVM shows a recall of 73% while RF shows a recall of 74%. Regarding precision, RF achieves 72%, and SVM reaches 73%.

The Receiver Operating Characteristics (ROC) [60] curves obtained from the

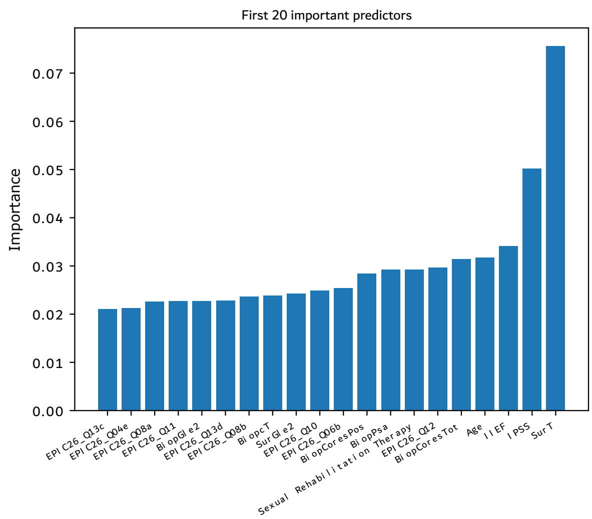

Figure 2.2. Importance of the different features in case of the binary classifier

two classifiers are displayed in Figure. 2.3. Classifiers with different operating conditions are shown in this graph as true positive rates versus true negative rates. This analysis is meant to choose the operating condition that represents the best trade-off between false positives and false negatives, depending on the purpose of the study. The dataset is split into two parts for training and testing, with 80 percent being used for training and 20 percent for testing. The AUC is also reported. This value is 76% for RF and 79% for SVM. Results are encouraging, suggesting that it is possible to predict whether patients will feel good or not with their sexual function after the surgery only by looking at the information available before the surgery. A study on a larger dataset would allow to confirm these findings.

Following the application of the ML classifiers and knowing their performance, a further analysis has been performed to identify which types of cases will improve performance. Conversely, it is important to take a closer look at the underlying causes of these results to understand why the classifier performance is not satisfying in some cases. In the following, we only consider SVM model in order to investigate the causes of the performance. In the figures that follow, orange and red colors are used to indicate the cases predicted incorrectly, while blue and green show the instances correctly classified. In particular, we report the cumulative distribution function (CDF) of the first five significant predictors. Figure. 2.4 shows the CDF of tumor stage (SutT). The figure shows that the

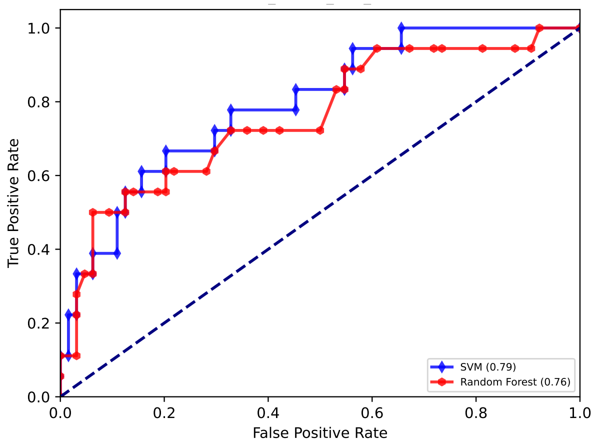

Figure 2.3. ROC curves of both used ML-models (AUC values)

highest number of incorrect predictions for class zero (orange), corresponding to a group without any problems but classified as having some problems, occurred when the tumor stage was eight. In contrast, patients with tumor levels of five had a relatively high probability of incorrect predictions for class one (red), which includes patients with any level of issue 12 months after surgery. As an interesting side note, most of the correct predictions were found for cases with tumor stage eight, while most of the correct predictions for class zero were found for cases with tumor stage five.

Figure. 2.5 shows the CDF plot of the the IPSS score. Class zero, i.e., the group with no problems indicated with green lines, the IPSS score around 0 provided the best prediction. For class one (reported in blue), 5-15 IPSS scores were best at predicting accurate and incorrect predictions.

In Figure. 2.6 the CDF of the IIEF score is reported. As shown, there is a tendency for ML classifiers to provide better performance with class zero (represented by the green line) when the score approaches zero. However, scores between 10 and 15 are associated with the most incorrect predictions in the same class. The blue color indicates that the classifiers perform better in the other class when the IIEF scores are between 15 and 20.

Figure. 2.7 shows the CDF of the age. Results obtained for class one (blue line) show that patients aged 65 or older who experienced a specific problem had the greatest accuracy in prediction. In class zero (orange line), a number of cases without any issues were predicted incorrectly regarding people in the age of 70.

Figure. 2.8 shows the CDF of the number of biopsy cores taken. Class zero is represented by the green line, which shows that one biopsy core can accurately

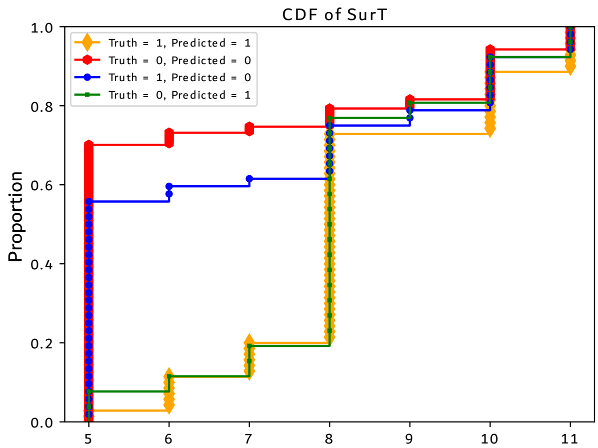

Figure 2.4. CDF of SurT

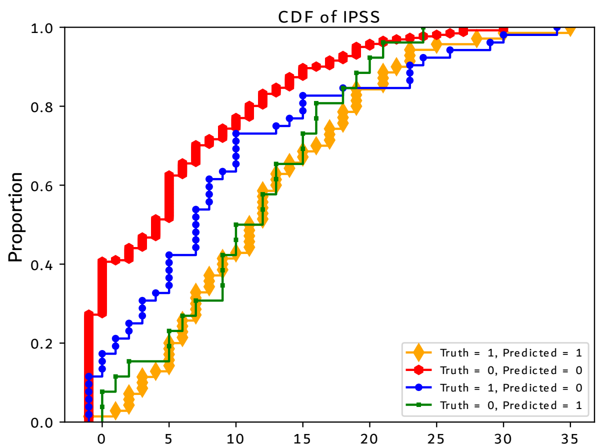

Figure 2.5. CDF plot of IPSS

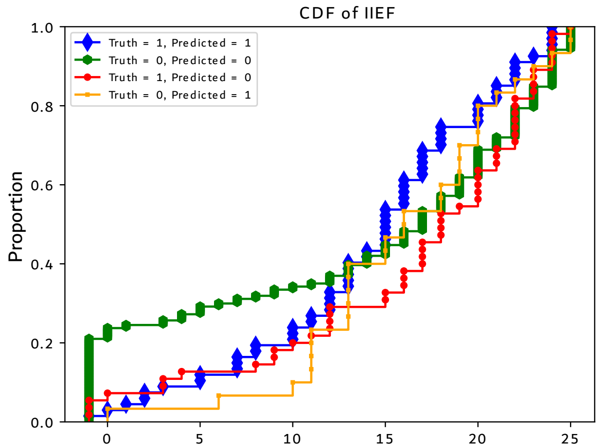

Figure 2.6. CDF of IIEF

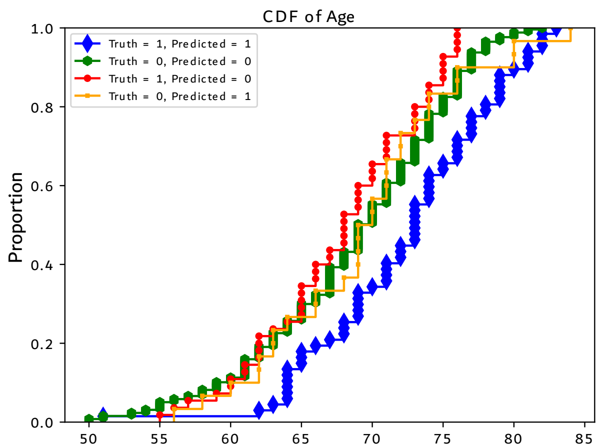

Figure 2.7. CDF plot of Age

predict more than 50% of the cases. For class one (represented by the blue line), most correct predictions were made with one biopsy core, but two biopsy cores still produced relatively accurate predictions.

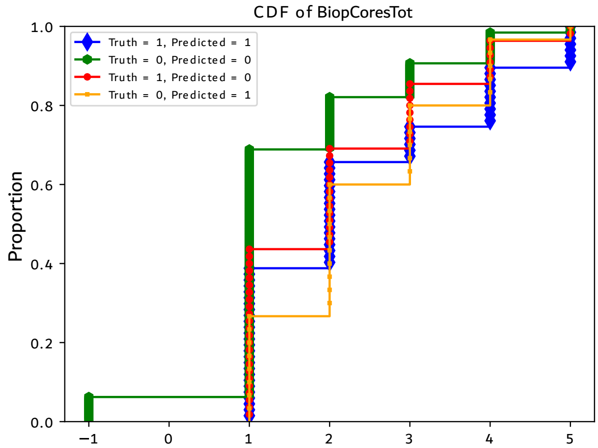

Figure 2.8. CDF plot of BiopCoresTot

2.4 Discussion

To our knowledge, no formal studies have been conducted to predict the sexual function of cancer patients following prostatectomy surgery. This knowledge gap may impede the development of effective interventions to address the issue, which is particularly important given the prevalence of this surgical intervention.

In our research, we gathered data on demographics, clinical information, and health-related quality of life by administering the EPIC-26 questionnaire. In the preliminary phase of this work, various ML algorithms were employed to address the assumed target, and the two best ML algorithm results are presented in this chapter, namely RF and SVM. In particular, the best performance was achieved when we used ML methods for binary classification and understanding whether the patients will experience any sexual function-related issues following prostatectomy surgery. We have used a meticulous methodology that employs two algorithms to process input data with utmost care and precision. The aim is to discern baseline variables that exhibit strong correlations with the sexual function of the patients in the future.

The methodology used in this study has been instrumental in determining the most predictive baseline features, which can help improve our understanding of the complex interplay between clinical and demographic variables. Based on our findings, the sexual function of the patients has assassination with a value of the group of variables such as patient age, IPSS, IIEF, SurT, and BiopCoresTot. The analysis indicates that a comprehensive understanding of a patient's future cannot be based on a single predictor. Instead, by taking into account a group of variables, it is possible to gain a deeper insight into the patient's prognosis.

We conducted a comparison between traditional statistical methods like Pearson correlations and logistic regressions with modern ML methods. Our analysis revealed that the traditional statistical methods Ire consistent with the ML ones regarding the most important predictors. The results of the logistic regression analysis indicated that the self-reported sexual life prediction was affected by the level of IPSS before surgery.

While recognizing the potential of the methodology described in this study to prognosticate and uncover previously unknown relationships betIen clinically significant variables, it is important to be cautious when interpreting the results presented, given the following limitations. Our heterogeneous data collection methods, which consisted of medical record reviews and phone interviews, may have impacted the accuracy of the results. In addition, the relatively small size of our dataset may not adequately represent real-world scenarios. It is imperative that these limitations be taken into account when interpreting our findings. Despite any limitations, we have confidence that the approach outlined in this study offers useful perspectives and may be employed in upcoming research to clarify the connections between significant factors. we are of the opinion that this model has the capacity to promote public health initiatives towards healthcare 4.0.

2.5 Conclusion

The purpose of this study is to understand if and how patient conditions after surgery can be predicted using information available before the execution of the surgery in order to gain insight into how AI methods can perform while facing real data. we concentrated in particular on the sexual life of patients as reported by themselves. Our case study utilized baseline and one-year demographic and clinical information of patients as the basis for our study. we used statistical and ML algorithms trained on various information about patients collected before surgery. The algorithms were able to estimate patients conditions using this information. The conditions estimated are the self-assessment of sexual function after surgery, determined using the EPIC-26 questionnaire. Our dataset consists of demographic information of 524 case studies and the results of their assessment of their HRQOL before and one year after surgery. In total, we had 57 characteristics for each patient. The results show that ML methods can not only predict the self-assessment of sexual function of patients one year after surgery but also identify important and critical factors such as age, IPSS, and IIEF that will influence patient sexual function in the future. Furthermore, we compared ML methods with well-established statistical methods such as Pearson correlations and logistic regressions. Based on the findings, traditional statistical methods were consistent with the ML ones with respect to the most important predictors. Results of logistic regression show that the prediction of self-reported sexual life is influenced by the level of IPSS before surgery. Results obtained with ML algorithms confirmed that IPSS before surgery is the most important feature (or predictor). Moreover, they also evidenced that the most recent Prostate-Specific Antigen (PSA) level (measured in ng/mL) in the blood and the age of the patient also have significant importance for this prediction. RF and SVM models were able to accurately predict sexual function after surgery with an accuracy of up to 71% when using multiclass classification. With binary classification, an accuracy of 74% could be achieved. we believe that this model has the potential to enhance public health initiatives in the direction of healthcare 4.0. In order to ensure that patients are prepared for what to expect after a procedure, healthcare providers can use this information as a tool to inform patients. Furthermore, the model can be used to provide a better understanding of the recovery outcome following surgery and to help them create better care plans based on those results. The ultimate goal of this project is to improve patient outcomes as much as possible.

Chapter 3

Mental Health

3.1 Introduction

The field of psychiatry has received a great deal of attention over the past few decades as one of the most important aspects of healthcare systems. In particular, the intersection of advances in therapeutic modalities of mental health and modern innovations contributed to the advancement of this field. The COVID-19 pandemic started an era along with fundamental changes in the lifestyle of humans and, subsequently, their mental health conditions. It is an undeniable fact that psychiatric symptoms among populations around the world have increased as a result of it. A diverse range of treatment methods exists that can be employed to reduce or eliminate psychiatric symptoms and improve mental health of individuals. Among these methods, tDCS has emerged as a promising approach for addressing mental health challenges, particularly those associated with depression.

The objective of this chapter is to scrutinize the predictive landscape surrounding mental health from two distinct yet interconnected perspectives. This examination aims to gain a comprehensive understanding of the predictive factors that influence mental health, evaluating the potential of AI methods to deal with the complex nature of mental health data and supporting psychiatrists in order to make a wise decision. Initially, we examined the impact of lockdowns such as COVID-19 on the mental health of different groups, including healthy individuals and those with mental health problems. Then, in the following work, the impact of tDCS treatment method on depressed individuals. We used HDRS tool to measure the level of depression in the participants. Both works aimed to decipher the crucial predictors that dictate the effect of lockdown and tDCS treatment on the mental health of attendees, as well as the potential of AI methods in this medical field.

This pioneering study aimed to examine AI-based methods in the mental health field and investigate the positive consequences of such techniques by navigating through predictive frameworks, patient characteristics, treatment modalities, and environmental influences.

3.2 Predicting the Severity of Lockdown-Induced Psychiatric Symptoms

3.2.1 Introduction

In the framework of a multi-modal conceptualization of mental health, it is well known that environmental stressors might, in most cases, elicit the onset of psychiatric diseases in vulnerable individuals or increase the severity of symptoms in psychiatric patients [88]. Starting in late 2019, the COVID-19 pandemic caused millions of deaths and severe physical complications in the global population. Indirect consequences of the health crisis, as well as social restrictions and economic constraints, led to psycho-social repercussions due to forced social distancing, disruption of stable behavioral patterns, and anxieties over the future. Indeed, several studies observed an increase in the incidence of psychiatric disorders in the general population [168, 135] even if contrasting results have also been reported. For example, a higher risk for developing severe depression and anxiety symptoms has been found in respondents of an online survey with a self-reported history of mental health problems [49], as well as in respondents directly tested for the restriction of physical activity [59]. Patients with pre-existing psychiatric conditions have been further affected by reduced access to psychiatry and psychotherapy services, which probably contributed to the exacerbation of their symptomatology [97]. In fact, it has been reported that, during the pandemic, psychiatric patients exhibited higher anxiety, depression, and stress levels with respect to the general population and showed a significant increase in the frequency of suicidal thoughts and episodes of insomnia [73, 80]. In this scenario, the early detection of the most vulnerable individuals among the general population and psychiatric patients would allow to prevent the onset or worsening of anxiety and depression symptoms, so reducing the burden of the pandemic on mental health and on national health systems. Several previous studies have used regression models to predict the evolution of psychiatric disorders. For example, Ready et al. used hierarchical regressions for predicting patients' future substance use, social behaviors, risky behaviors, and psychological distress [130]. Furthermore, generalized linear regression was used for investigating the personality traits associated with depressive symptoms in the general population across a fifteen-year period, and it has been proven that personality traits were poor predictors of depression for specific time points [134]. Similarly, Wang et al. [170] used an operating characteristic curve and logistic regression analyses and compared the accuracy of clinicians with the accuracy of regression on a standardized scale in terms of evaluating the risk of future suicide attempts in a sample of psychiatric residents. A recent investigation [18] applied multiple logistic regression analyses to identify the socio-demographic and clinical factors associated with depression in a large sample of patients in the context of a first-episode psychosis program. The same method was used for identifying the predictors of depression and anxiety in Brazil during the initial outbreak of COVID-19, and it has been observed that females, younger adults, and individuals with fewer children had a higher likelihood of developing depression and anxiety symptoms [177]. During the past two decades, ML [98, 101, 100] has become one of the most well-known methods used for several purposes, including prediction.

In this case, we examined the ML approach to determine if ML approach can identify the demographic and clinical characteristics of individuals with a higher risk of a poor psychological outcome during the COVID-19 pandemic-related lockdown. Through this case study, we gained insights into the reliability of AI methods for identifying critical predictors and the performance of ML in unusual circumstances, making it suitable for a wide range of situations. We tested this approach on a sample of healthy individuals and two different populations of psychiatric patients, i.e., patients with Obsessive Compulsive Disorder (OCD) [160] and patients with Adjustment Disorder (AD) [121], whose demographic and clinical information was collected before and during the Italian lockdown. Although the present work aimed not to conclude specific clinical issues, we selected these two psychiatric populations because we hypothesized that they could have been affected by the lockdown more than others. In fact, on the one hand, OCD patients often suffer from obsessive thoughts of contamination, which the risk of COVID-19 contagion could have exacerbated. In addition, compulsive washing and cleaning, which are also typical OCD symptoms, could have been worsened by the continuous invitation of hand-washing and hygiene coming from governments and health authorities [66]. On the other hand, patients with AD are defined as having developed emotional or behavioral symptoms out of proportion to the severity or intensity of an identifiable stressor [121]. We hypothesized that this could imply a greater vulnerability to the psychological impact of the lockdown. Since Italy was the first European country hit by the pandemic and the first to enact large-scale lockdown measures, people's perception of the restrictive measures was probably more dramatic than anywhere else. Starting from this unique observatory, this study aimed to implement a tool that is able to provide prognostic elements that could support clinical decisions during large-scale lockdowns. In particular, the practical application of this tool would be the development of a risk-scoring protocol used on a routine basis by clinicians and service providers for the primary and secondary prevention of psychiatric symptoms.

The contributions of this work can be summarised as follows: (i) we created a methodology based on supervised machine learning to identify predictors of the severity of psychiatric symptoms during the Italian lockdown as well as to predict the degree of severity to be used for building healthcare 4.0; (ii) we applied this methodology to a use case on a small sample of individuals showing how it can be used in a real case; and (iii) we presented preliminary results showing that our models are able to predict depression, anxiety, and obsessive-compulsive symptoms during the lockdown with up to 92% accuracy based on demographic and clinical characteristics collected before the pandemic.

3.2.2 Materials and Methods

Participants

We enrolled two convenience clinical samples (one involving OCD patients and one involving AD patients) and one sample of healthy subjects. All patients were being treated at the University Hospital Federico II of Naples (Italy). Eligibility criteria for patients were 18-70 years of age and psychopathological stability at baseline evaluation. Eligible healthy subjects were 18-70 years of age without a history of psychiatric illness. They were selected in order to obtain mean values of their demographic variables (i.e., age, sex, education) comparable to those of clinical samples. Exclusion criteria were having COVID-19 (even suspected); being exempt from quarantine for any reason, being hospitalized or having been hospitalized during the lockdown, and having a severe chronic medical disorder.

The demographic characteristics of participants are reported in the supplementary material.

All procedures were approved by our institutional review board and in accordance with the Declaration of Helsinki and its later amendments. All participants provided their informed consent.

Clinical Measures and Data Collection

Demographic data (e.g., age, sex, education, profession) and information about medical history (e.g., psychiatric diagnosis, family history of psychiatric diseases, medical comorbidities) were collected. Moreover, the scores of two to four psychiatric scales were gathered from medical records or through telephone calls, depending on the group.

The Yale-Brown Obsessive-Compulsive Scale (Y-BOCS), a clinician-administered rating scale, assesses the current and lifetime presence of OCD symptoms and their severity the week before evaluation. It consists of a symptoms checklist with 54 obsessions and compulsions dichotomously scored as present or absent and of a rating scale assessing the severity of the current symptoms in terms of time spent, interference, distress, resistance, and control. The rating scales comprise ten items: five for obsessions and five for compulsions. All items have a Likert-type scale ranging from 0 to 4 so that it is possible to obtain a total score of 0-40 for the overall obsessive-compulsive symptoms and two subtotal scores of 0-20 for obsessions and compulsions separately [61].

The Brown Assessment of Belief Scale (BABS) is a seven-item clinician-administered semi-structured scale designed to assess the degree of conviction and insight that patients have concerning their beliefs. It consists of seven items: the first six items were added to obtain the total BABS score, while an additional item (ideas of reference) is not included in the total score. Each item is rated from 0 (non-delusional, or the least pathological) to four (delusional, or the most pathological) [45].

The Beck Depression Inventory-II (BDI-II) is a multiple-choice self-report inventory that consists of 21 items assessing the affective, cognitive, and physical symptoms of depression. Each item is rated from 0 to 3 (from the least to the most severe) [20]. The State-Trait Anxiety Inventory-Y (STAI-Y) is a commonly used measure of trait and state anxiety. Form Y, its most popular version, has 20 items for assessing trait anxiety and 20 for state anxiety. All items are rated on a 4-point scale (e.g., from "‘Almost Never" to "Almost Always"). Higher scores indicate greater anxiety [155].

Figure. 3.1 shows the cumulative distribution function of all scales. Note that BDI-II and STAI-Y are related to all three groups, whereas the Y-BOCS and the BABS are only related to the OCD patients because they are not meaningful for non-OCD individuals.

In line with the predictive purpose of this study, all scales were administered at two different time points, i.e., before the pandemic (baseline) and during the Italian lockdown period (follow-up), namely from March to June 2020. The baseline scores of the OCD and AD samples were gathered from the patients' clinical records (STAI-Y and BDI-II for both groups, Y-BOCS and BABS for the OCD group) since the aforementioned clinical scales are collected on a routine basis. Given that patients could not be reached in person during the lockdown, the follow-up assessment was performed through a telephone call by an expert examiner blinded with respect to the patient's diagnosis. For healthy subjects, STAI-Y and BDI-II scores at both time points were collected during the same telephone call, with the baseline evaluation being retrospective in nature. The methodology worked on the values of these four scales administered before the lockdown and used such values as features (i.e., input values) for the prediction. The predicted value was then one of the rates of the same scales but during the

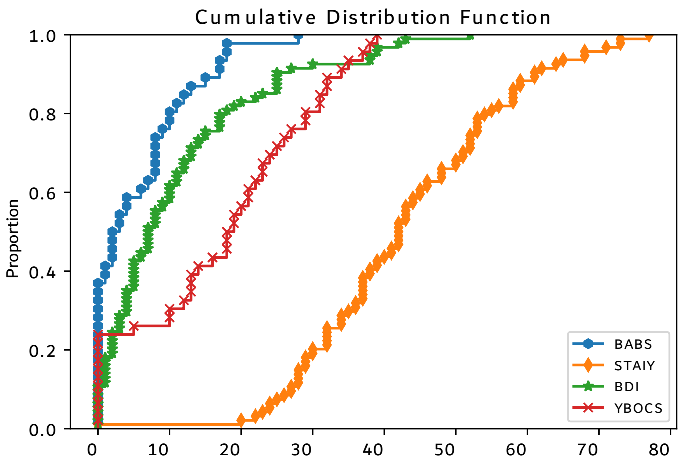

Figure 3.1. Cumulative distribution function of all scales at T0.

lockdown. In conclusion, looking at the values of the scales before the lockdown together with other information regarding the participants (i.e., demographic data and information about medical history), properly trained ML algorithms can predict the values of any scale during a lockdown.

Data Pre-Processing

In order to have uniform data, the dataset was cleaned, pre-processed, and normalized. After gathering all the data, the scores of each questionnaire were calculated and added to the database. Thereafter, in order to make the prediction, the calculated scores were clustered into three to five groups. For the BABS scale, the scores were categorized into three groups: 0-5, 6-15, and >15. For the Y-BOCS, the categorisation by [103] was used: <7 likely to be sub-clinical, 8-15 mild OCD, 16-23 moderate OCD, 24-31 severe OCD, and 32-40 extreme OCD. The category used for BDI-II was: 0-13 for minimal depression, 14-19 for mild depression, 20-28 for moderate depression, and 29-63 for severe depression [20].About the STAI-Y scale, the possible categories were "no or low anxiety" (20-37), "moderate anxiety" (38-44), or "high anxiety" (45-80) [86].

On one hand, it is important to underline that the individual and not the grouped scores were used as input for the prediction. On the other hand, the clustered values were used as the values to be predicted. For example, when classifiers predicted the BDI-II, the category of the scale of BDI-II as previously defined was used, and when the classifier predicted "mild depression," it meant that the score was somewhere between 20 and 28.

3.2.3 A Preliminary Study on a Real Dataset

This section shows how the methodology can be applied in a preliminary study. We highlight that the results are to be considered preliminary as the dataset is related to a small number of subjects.

Data Used

We used data from 94 subjects assessed at two different time points, i.e., at baseline (before the pandemic) and at follow-up (during the COVID-19 lockdown). Of these subjects, 46 were OCD patients, 19 were AD patients, and 29 were healthy subjects. The baseline demographic and clinical characteristics of the three groups are reported in the supplementary materials.

Results Obtained

Four different case studies were performed to study the impact of the COVID-19 lockdown and any other kinds of lockdown on psycho-pathological patients and healthy individuals. In each of these analyses, the objective was to predict the status of target subjects measured through one of the scales during the first lockdown of COVID-19. Moreover, for each scale to be predicted, we evaluated the importance of each feature for prediction in order to capture the variables (demographic, medical history, and psychiatric symptoms before the lockdown) that play a major role in the emergence of psychiatric symptoms during the lockdown.

- **Predicting Depression Symptoms**

  In the first case study, the depression symptoms (measured through the BDI-II scale) during the first lockdown (T1) were predicted based on data gathered before such lockdown (T0). Figure 3.2 shows the association between BDI-II at time T1 and all other features at baseline (T0). This figure shows that the answers to the BDI-II and STAI-Y scales at time T0 played an important role with respect to other features.

  Table 3.1 reports the performance measures of the prediction computed as described previously. The results are obtained through a 10-fold cross-validation. The table shows that the classifiers predicted with accuracy up to approximately 70% percent and F1-score up to approximately 63%. Precision and recall range between 55% and 68%. Furthermore, the table shows that the performance of both classifiers for predicting BDI-II at time

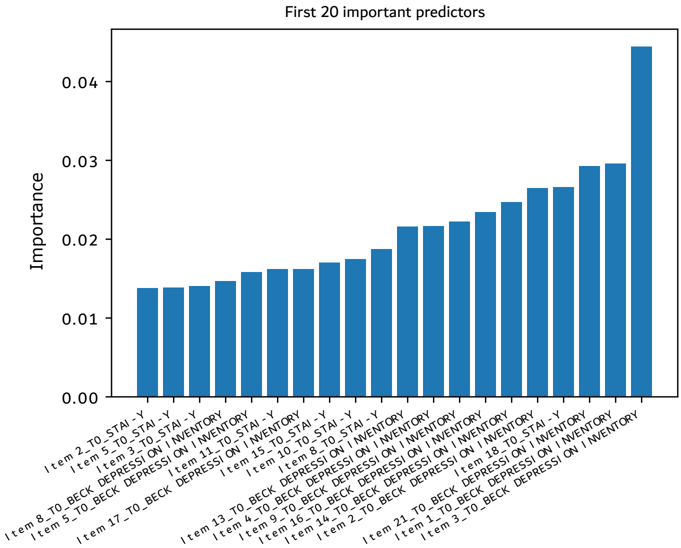

Figure 3.2. Importance of the different features for predicting depression symptoms.

T1 was approximately the same. The RF classifier shows slightly better performance.

Table 3.1. Results of the prediction of depression symptoms.

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.60	0.68	0.63	0.68
SVM	0.56	0.66	0.59	0.66

- **predicting Anxiety Symptoms**

  The second case study was aimed at predicting the anxiety symptoms (measured through scale STAI-Y) at time T1 by using the rest of the features of the dataset, including demographic data and the answers to the questionnaires of BABS, Y-BOCS, STAI-Y, and BDI-II at time T0. Figure 3.3

shows that the first ten most important features are a combination of different items of BDI-II and STAI-Y at time T0, as well as age which also plays an important role.

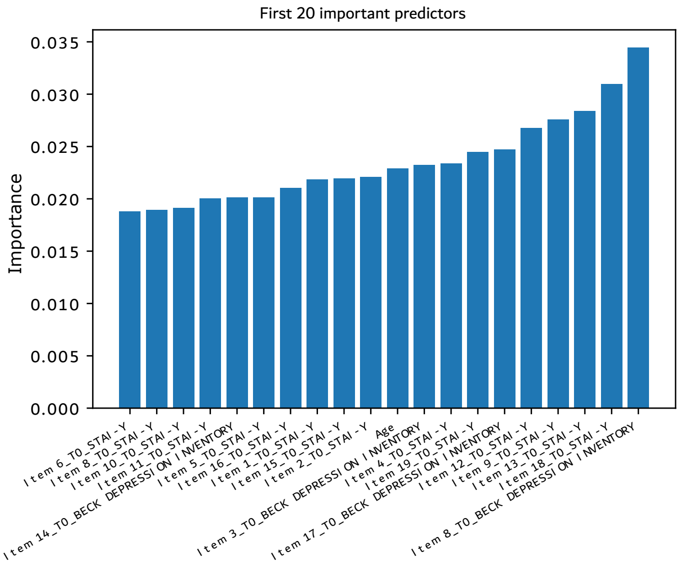

Figure 3.3. Importance of the different features for predicting anxiety symptoms.

Table 3.2 shows the performance of classifiers for predicting STAI-Y at time T1. As for the previous case study (predicting BDI-II at time T1), RF obtained a prediction performance higher than the SVM. The accuracy is up to 75% and F1-score is up to 70%.

- **Predicting Obsessive and Compulsive Symptoms**

  In the third case study, the obsessive and compulsive symptoms (measured through scale Y-BOCS) was the class to be predicted. Figure 3.4 shows the role of each feature in the prediction process. According to this figure, selected items of the scale Y-BOCS at time T0 are the most important ones.

  According to Table 3.3, the performance of ML classifiers for predicting the condition of subjects in terms of Y-BOCS values are approximately the same for both classifiers. The accuracy is up to 71%, and the F1-score is up to 67%. The difference is mainly related to recall and accuracy. For

Table 3.2. Results of the prediction of anxiety symptoms.

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.68	0.75	0.70	0.75
SVM	0.61	0.67	0.63	0.67

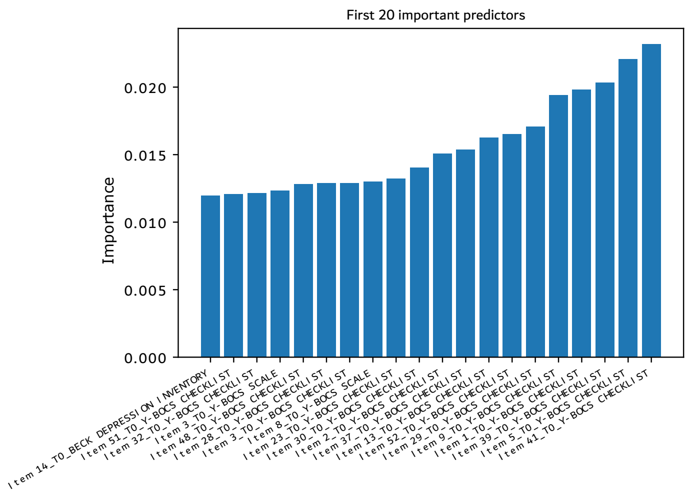

Figure 3.4. Importance of the different features for predicting obsessive and compulsive symptoms.

RF, it is approximately 70%, while for SVM, it is approximately 71%. Precision and F1-score are 65% and 67%, respectively.

Table 3.3. Results of the prediction of obsessive and compulsive symptoms.

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.65	0.70	0.67	0.70
SVM	0.65	0.71	0.67	0.71

- **Predicting Belief Symptoms**

  In the last case study, belief symptoms at time T1 (evaluated through the BABS scale) is the class to be predicted. Figure 3.5 shows that there is a strong correlation between the scale BABS at time T1 with scales BABS and Y-BOCS at T0. These scales have a remarkable impact on the results in this case study.

  Figure 3.4 shows the performance of the two considered classifiers for predicting the scale BABS at time T1. The classifiers achieve accuracy values up to 92% and F1-score values of up to 90%. Comparing the two classifiers, precision values are mostly similar for both classifiers, while recall values are higher for the SVM.

Table 3.4. Results of the prediction of belief symptoms.

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.90	0.90	0.89	0.90
SVM	0.90	0.92	0.90	0.92

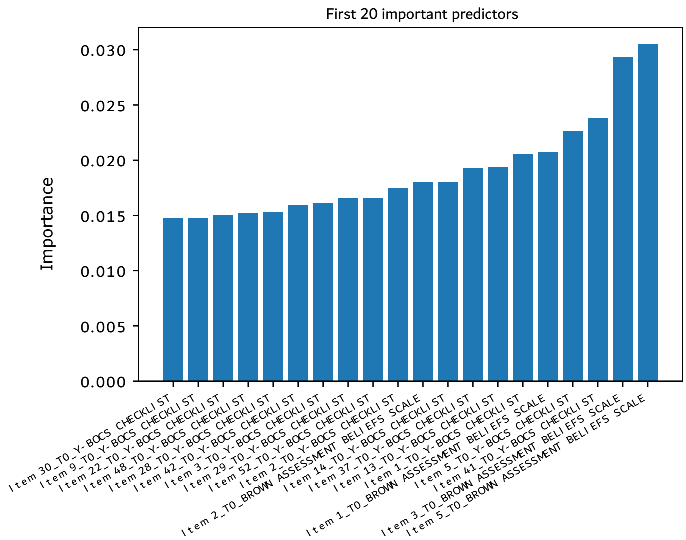

Figure 3.5. Importance of the different features for predicting belief symptoms.

3.2.4 Discussion

A few studies have used ML to predict the consequences of the COVID-19 lockdown in terms of psychiatric symptoms, and to the best of our knowledge, none were based in Italy. Italy was the first European country to be hit by the COVID-19 pandemic and the first to enact an extended large-scale lockdown. Moreover, single regions such as Campania implemented additional restrictive provisions. This was an unexpected and unprecedented contingency, provoking a great amount of uncertainty about the future, psychological distress, worry, and anxiety in the population. In fact, nationwide lockdowns and stay-at-home restrictions had never occurred in recent history in Europe before March 2020.

In the present study, we collected the demographic and clinical characteristics of healthy individuals and psychiatric patients from the Italian region of Campania before and during the COVID-19 lockdown with the aim of identifying possible predictors of the severity of psychiatric symptoms. We tackled this issue with two ML algorithms, i.e., RF and the SVM. In the preliminary phase of this study, we also tested other algorithms, including kNN and Naive Bayes, but RF and SVM achieved the best performance. We set up a methodology based on these two algorithms that carefully processed the input data in order to identify which baseline variables were strongly associated with the clinical evolution at the follow-up.

Our models allowed us to identify the most predictive baseline features, therefore possibly increasing the knowledge about the interplay between specific clinical variables and/or between clinical and demographic variables. More importantly, our results showed that the severity of psychiatric symptoms such as depression, anxiety, and OCD during the lockdown could be predicted by taking into account items of the BDI-II, STAI-Y, and Y-BOCS collected at the baseline and no single predictor can be used for this task. In particular, the severity of anxiety symptoms, expressed as a total score on the STAI-Y collected during the lockdown, was best predicted by the score on some STAI-Y items and the score on a few BDI-II items gathered at the baseline. Similarly, the five most important features for predicting the BDI-II total score during the lockdown were the BDI-II items 3, 1, 21, 2, and item 18 of STAI-Y collected at the baseline.

Despite the prognostic potential of the described methodology and the opportunity to unravel unknown relations between clinically meaningful variables, the above results should be considered cautiously because of the following limitations. First, we employed heterogeneous methods (i.e., medical records vs. phone calls) to collect baseline data on psychiatric patients and healthy subjects. This difference makes it difficult to compare baseline findings in the three groups. Moreover, baseline and follow-up data were collected for healthy individuals within a single telephone call during the lockdown. This might have affected pre-pandemic data reliability due to physiological oblivion as well as state-dependent memory recall and the selective attention biases of respondents. Finally, since it is not possible to have baseline clinical information about healthy subjects (non-clinical by definition), we acknowledge that with respect to this set of data, the application of our methodology is just a hypothetical example of how to use baseline information to predict the psychiatric outcome of a given population during a lockdown. However, on the one hand, for healthy subjects, it is still possible to use demographic information that is always available, apply the described method, and identify the most vulnerable subjects among the general population, which is of interest to public health policies. On the other hand, the baseline clinical information is already available for clinical populations and can be combined with demographic characteristics, so yielding potentially sensitive and helpful predictors of psychiatric outcomes, which can be used to set up prevention policies in psychiatric patients. In this view, our study might pave the way for future investigations aiming to practically apply the described models for the implementation of a risk-scoring system. Such a system would allow clinicians and service providers involved in psychiatric prevention to enter the most appropriate indicators for the determination of the risk profile of a given individual into a computer and consequently enact personalized and well-timed therapeutical interventions. In our specific case of lockdown-induced psychiatric symptoms, these early interventions for at-risk subjects could consist of programs of more frequent telehealth psychiatric evaluations and/or psychotherapy sessions for those already in treatment and the implementation of regular sessions of psychological support for healthy subjects.

The main purpose of our data analysis was to show how to apply the methodology to a real sample practically in line with healthcare 4.0 and not to draw clinical conclusions. As a matter of fact, in the case of providing a larger sample for the algorithm, it would be possible to not only give more conclusive results but also explore the cases for which the algorithm provides accurate results and the cases for which the prediction is not accurate. Moreover, future studies could use this methodology to explore the predictive value of a greater number of variables with respect to those we considered, e.g., biochemical or imaging markers, which would increase the accuracy of the prediction.

3.2.5 Conclusions

In conclusion, here we propose an ML approach to identify the demographic and clinical variables suitable to predict the onset and/or worsening of psychiatric symptoms during a dramatic large-scale event such as the COVID-19 lockdown. In the present work, we used the demographic and clinical information of two psychiatric populations (i.e., OCD and AD) and healthy individuals as case studies and demonstrated that the RF and SVM models can predict the severity of depression, anxiety, and obsessive-compulsive disorder symptoms with an accuracy of up to 92% in each population. Should our preliminary findings be confirmed by future studies on larger samples, this model could be applied in cases of dramatic events causing large-scale lockdowns in two possible ways: in clinical settings, for prevention of symptom deterioration in higher-risk psychiatric patients, by health authorities, in order to implement public health programs such as clinical AI-DSS specifically targeting vulnerable clinical and non-clinical populations.

3.3 Predicting Hamilton Depression Rating After Transcranial Direct Current Stimulation Treatment and Detecting Important Predictors

3.3.1 Introduction

The tDCS is a non-invasive technique for stimulating the brain by delivering weak direct currents to the brain through sponge electrodes to stimulate brain function [123]. This led to the use of tDCS to treat major depression. Structural and functional changes are seen in several cortex regions associated with depression, such as the left dorsolateral cortex and ventromedial cortex, as well as the amygdala and hippocampal area of the brain [123]. A simple model for the function of tDCS involves shifting resting membrane potentials of neurons towards depolarization with anodal stimulation and towards hyperpolarization with cathodal stimulation. To achieve the desired outcome, tDCS is used to either increase or decrease neuronal firing rates, depending on the polarity of the electrode used. It is believed that tDCS has a neuromodulatory effect, which is why it is often used for this purpose [123].

In antidepressant clinical trials, the HDRS has been commonly considered for evaluating depression severity, treatment efficacy, and assessing antidepressant efficacy [76]. The depression scale was first published in 1960 under the name HDRS. HDRS has been used for years to help determine the severity of depressive states and the responses of patients to treatments, as well as helping diagnose depression in patients [33]. The HDRS is a limited instrument that only focuses on a few clinical symptoms and does not assess positive emotions [56].

Due to the availability of a large amount of data in the healthcare sector, AI has become extremely popular. It is a method for analyzing patient data to extract useful information. Significant advances have been made in computer vision, natural language processing, and automatic speech recognition due to progress in this field [58]. In the paper led by Ghandeharioun, the authors conducted a study where they developed and tested the effectiveness of ML methods for predicting the HDRS based on the objective data collected passively and continuously by E4 wearable wristbands and sensors attached to Android smartphones [57]. Based on the findings of the study, it has been found that poor mental health is associated with unregular sleep, reduced motion, fewer incoming messages, a decrease in the variability of location patterns, as well as increased asymmetry of electrodermal activity between the right and left wrists [57]. In this field, a study evaluated three ML classifiers to predict treatment outcomes for late-life depression by using sociodemographic characteristics, baseline clinical self-reports, cognitive tests, and structural magnetic resonance imaging (MRI) features. [64]. The research, led by Benoit et al., developed a model that can effectively predict remission status in patients who have undergone desvenlafaxine treatment. The study analyzed 3,399 individuals who were diagnosed with major depressive disorder using machine learning algorithms. The findings revealed that the linear support vector machine had an accuracy rate of 69.0% in the holdout set [22]. According to another study published in this field, the researchers investigated whether certain factors identified by structural magnetic resonance imaging (MRI) techniques can be used to predict the response of a person to electroconvulsive therapy (ECT). It has been found that binary classification has been a successful tool for predicting the effectiveness of ECT and that support vector regression has been found to be a significant predictor of the relative reduction in HDRS score following ECT [132]. As a result of a project led by Tazawa, wearable device data was utilized in developing a ML algorithm to screen for depression. Based on the results of the study, wearable devices and ML have the potential to be useful not only for identifying depression but also for assessing the severity of depression [163]. It has been demonstrated in another pilot study that ML and MRI-derived electric field models can be used as proof of concept for precision cognitive interventions aimed at improving working memory. An analysis of MRI-derived tDCS current models was performed using pattern recognition techniques to identify individual prognostic factors of tDCS treatment response with a precision of 86 percent [8]. In a pilot randomized controlled trial led by Marotta, the aim was to assess whether and how tDCS could improve balance and gait in patients with Multiple Sclerosis by using a method known as ML. There has been a demonstration in this pilot randomized controlled trial that tDCS may result in a non-sustained improvement in gait and balance among patients with MS. Using ML in such a case could suggest that the benefits could last for a long time [107].

In order to predict whether tDCS treatment for major depressive disorder is effective, the purpose of this study is to use supervised machine learning to identify predictors associated with the outcome and make predictions of HDRS. As well as the mentioned goal, we want to evaluate the potential of the presented methodology to be employed for supporting clinical decisions and providing better healthcare services. Using the methodology, we were able to apply it to a use case on a small sample of individuals, showing how it can be used in real cases. As a result of these findings, it can be concluded that the supervised machine learning method is able to predict HDRS after tDCS treatment based on the characteristics of the patients collected before the treatment. Furthermore, this study provided insight into the factors responsible for accurate predictions in the first place by using the methodology. Based on these results, we believe that further studies on this topic are necessary and that they will be encouraging to provide better care to the patients.

3.3.2 Materials and Methods

Participants and Clinical Measures

Data collection was conducted by contacting three clinical centers. An analysis of 169 subjects was conducted to examine how patients responded to tDCS treatment. Initially, all patients were experiencing a major depressive episode, either unipolar or bipolar, without psychotic features or suicide risk, and all were treated with tDCS. The tDCS treatments were carried out between 2013 and 2017 in one inpatient psychiatric clinic and two outpatient psychiatric clinics. Each patient underwent ten to twenty tDCS sessions on a daily or twice daily basis. During each session, a 1.5 to 2mA electric current was applied in correspondence of the patients' frontal lobe(s) for 20 to 30 minutes.

The used dataset contains data regarding various aspects such as the patient's gender, age, illness features (unipolar vs. bipolar depression) and duration, presence of comorbid psychiatric conditions, ongoing pharmacological treatment, tDCS protocol, and other relevant factors. As part of this dataset, HDRS ratings were also collected for pre- and post-treatment, which is one of the most widely used ratings for depression severity [172]. Table 3.5 provided a summary of the main characteristics of the participants. The table has three sections. The first section displays the age range of participants, the second section shows the portion of males and females, and the last section indicates the number of patients diagnosed with unipolar and bipolar disorders.

Table 3.5. Participants demographic charecters

Age	Mean	44.47
	Median	44.47
	Variance	519.81
	SD	22.80
	Q1	36.00
	Q3	61.00
Sex	Female	84.00
	Male	59.00
Diagnosis	Unipolar	105.00
	Bipolar	38.00

Data Pre-Processing and Analysing

Upon preprocessing the gathered dataset, we applied ML methods to demonstrate the impact of tDCS treatment on HDRS after treatment. It is used to determine the best ML classifiers for predicting the Hamilton depression scale as well as identifying impactful predictors on HDRS of patients after ten treatment sessions. In all analyses, a dummy classifier was used as a baseline for comparing the performance of different classifiers.

Table 3.6 shows the method, the label, and the dataset that was used in order to achieve the results. Two labels were defined, corresponding to what was described in the table: the first was defined as (T2-T0)/T0, while the second was defined as T2. The T0 refers to the HDRS at baseline, and the T2 refers to the HDRS ten sessions after the tDCS therapy. The proportion of the score reduction is considered because it is the main indicator of response to an antidepressant treatment. In particular, a partial response is referred to when there is improvement in the initial phase of therapy, but symptoms continue to be present. This corresponds to a decrease in starting scores on the HDRS scale between 50% and 25%. A full response to the treatment is when a 50% score reduction occurs. Moreover, the precise value of the T2 is considered to evaluate the ML methods capability for predicting T2. As well, two methods (binary classification and multiclass classification) of ML were used. For multiclass classification, the target variables were categorized into three groups, and for the binary classification method, the target variables were divided into two groups. Furthermore, in order to predict the outcome of HDRS after tDCS treatment, two groups of variables were selected to represent the dataset (DB.v1 and DB.v2). DB.v1 is the first version of the dataset, containing demographic data and information regarding the medications used. The second version of the dataset included not only all variables of the initial dataset but also the combination of the medication taken and the code of the treating physician. we used the code of the treating physician because the center where the treatment has been performed (each doctor's code corresponds to a different center in a different part of Italy) could have affected the treatment outcome for different reasons, such as selection bias for facilities availability, treatment setting, technicians proficiency. Moreover, the reason for using the combination of medicines was that the concurrent administration of psychotropic drugs could affect the tDCS effects in a not predictable way.

It is important to stress the fact that data from target variables were not used to predict a HDRS after tDCS treatment.

Classifiers were evaluated according to four metrics: precision, recall, F1-score, and accuracy.

3.3.3 Obtained Results

An analysis of the obtained results using the described methodology is provided in the following sections.

Table 3.6. Followed methods to achieve the results

Label	Dataset	ML type
(T2-T0)/T0	DB.v1	Multiclass classification
		Binary classification
	DB.v2	Multiclass classification
		Binary classification
T2	DB.v1	Multiclass classification
		Binary classification
	DB.v2	Multiclass classification
		Binary classification

Feature Importance

This step identified the important predictors that were critical in predicting target labels.

Initially, the database version DB.v1 was used. It means that only demographic data and medicine-related data were available in the used database for the first label ((T2-T0)/T0). A key variable was identified from the analysis of the used database. Figure 3.6 shows the contribution each predictor made to the prediction of target values in both the binary and multiclass classifications. Figure demonstrates that HDRS (HAM-D 21 T0 tot) before treatment, along with the patient's age, diagnosis, the duration of the disease, as well as the gender of the patient, were the first five important predictors of outcomes. As compared with any other factors, the HDRS and the age of the participants had the most significant impact on the results.

In the next step, the items of the second database (DB.v2) were analyzed in order to establish exactly what role the predictors played in determining the result of the analysis as a whole. The second database, besides the demographic data and the used medications, also contains information about the doctor and the combination of prescribed medicines based on their category. Figure 3.7 displays the impact of each of the predictive variables in predicting the target label. According to the experiment results, the three most important factors that have the greatest effect on the prediction of the target label are the same as those considered in the previous analysis. The results of the current study showed that although all of the predictors had some level of impact on predicting target value, the combination of medicines taken was not as critical as other variables, like the primary diagnosis or information related to the treating doctor.

In order to understand the role that predictors played in predicting T2 as the target value, we evaluated their importance as part of the next group of

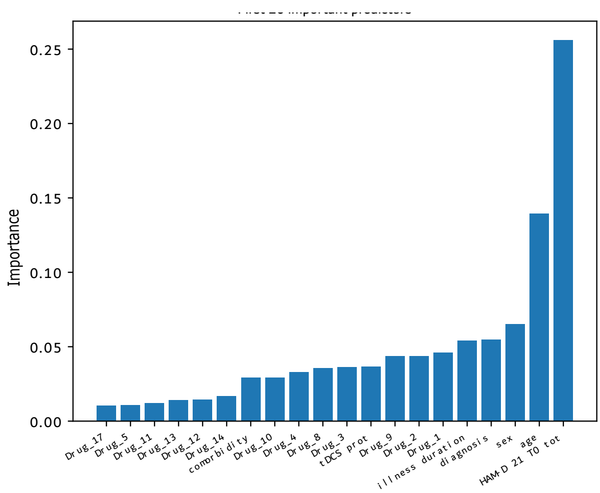

Figure 3.6. Importance of the feature for predicting (T2-T0)/(T0),DB.v1

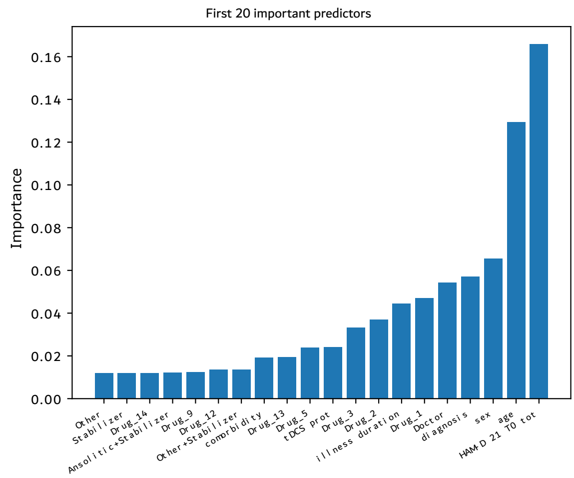

Figure 3.7. Importance of the feature for predicting (T2-T0)/(T0),DB.v2

analyses. For this assessment, only the variable from the first version of the dataset (DB.v1) was used. Figure 3.8 depicts the results of the aforementioned analyses, the HDRS at the baseline, and the age of the patients located in the first two places, followed by the variables such as gender, diagnosis, and length of illness located in the next places. Furthermore, the study revealed that when comparing the medicine that a person has taken and the duration of the illness, the length of the illness will be a more significant factor in determining how accurate the prediction will be than the medicine taken.

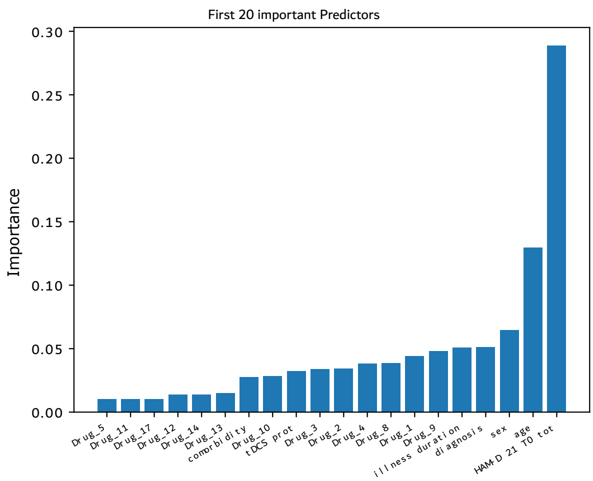

Figure 3.8. Importance of the features for predicting T2,DB.v1

Finally, a variable from the second version of the dataset (DB.v2) was used in the analysis, which contains the treating doctor and the combination of medicines based on their category, along with the first version of the dataset. Figure 3.9 illustrates the fact that had been shown in all analyses associated with detecting the important feature of the data. The results of this evaluation indicate that the HDRS before treatment, age, and gender of the participants were among the significant predictors of the target labels related to the condition of patients following tDCS treatment.

It should be mentioned that based on the results obtained during the analysis for both binary and multiclass classification, the important predictors were approximately the same.

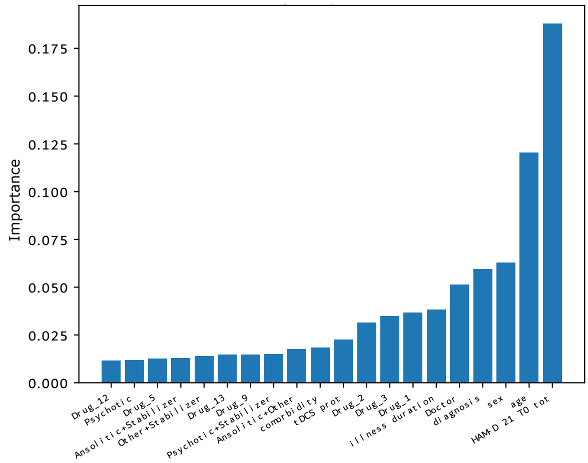

Figure 3.9. Importance of the features for predicting T2, DB.v2

Evaluating Machine Learning Algorithms

- **Predicting (T2-T0)/(T0), Multiclass Classification**

It has been determined that (T2-T0)/(T0) represents the label to be predicted in this analysis step. This was done as a first step by dividing the value of (T2-T0)/(T0) into three categories. Owing to the limited size of the utilized dataset, a methodology based on cut-off values was employed to optimize the performance of the ML model while also aligning with the clinical criteria for evaluating patient response to treatment. This approach was deemed necessary to ensure the accuracy and reliability of the ML model's predictions. By implementing this method, the model was able to effectively address the clinical objectives and produce results that were consistent with the desired outcomes. There were cases in the first two groups(> -0.12, and $-0.35 << -0.11$) who did not respond fully to treatment, whereas in the last group(< -0.34), they did respond almost fully.

Table 3.7 details the distribution described previously. In this study, one of the purposes was to gain a deeper understanding of how ML algorithms react when dealing with small-sized data sets. Table 3.8 shows the performance of eleven different ML algorithms, when they were empolied to predict (T2-T0)/(T0). The table shows that the total performance of all of the ML algorithms was less than fifty percent. In addition, it shows that the performance of all ML algorithms was comparable with the performance of a dummy classifier. Moreover, even though the performances of each individual ML algorithm

Table 3.7. Distribution of (T2-T0)/(T0) for multiclass

>-0.12	-0.35< <-0.11	<-0.34
61	51	57

differed, the result showed that all of them produced roughly the same results.

Table 3.8. Performance of the ML methods for predicting (T2-T0)/(T0), DB.v1

	PRECISION	RECALL	F1_SCORE	ACCURACY
Random Forest	0.38	0.36	0.36	0.36
SVM	0.27	0.31	0.27	0.31
K Nearest Neighbors	0.34	0.33	0.31	0.33
Decision Tree	0.37	0.37	0.37	0.37
Logestic Regression	0.31	0.31	0.30	0.31
MLP	0.33	0.33	0.32	0.33
Stochastic Gradient Descent	0.28	0.31	0.26	0.31
Adaptive Boosting	0.35	0.35	0.34	0.35
GradientBoosting	0.36	0.35	0.35	0.35
XGradientBoosting	0.40	0.36	0.37	0.36
Dummy Classifier	0.32	0.32	0.32	0.32

Table 3.9 depicts the confusion matrix of one of the most effective ML algorithms. Due to the similar performance of all ML methods, we only provided one of the confusion matrices in order to show how the methods worked when dealing with small datasets.

To investigate the capability of ML methods to predict (T2-T0)/(T0), DB.v2 was used to perform another set of analyses. This means that the combination of the drugs that the participants took and the code for the doctor who treated them have been added to the database. Table 3.10 illustrates the performance of all ML classifiers. The table indicates a slight improvement in predicting target value when using a new database.

Table 3.11 displays the confusion matrix when ML methods try to predict a target value. According to the table, the best classifier algorithm predicts most cases in the major class in order to provide better results.

- **Predicting (T2-T0)/(T0), Binary Classification**

Table 3.9. Confusion matrix of Multi-class classification, (T2-T0)/(T0), DB.v1

Group_one	Group_Two	Group_Three
7	1	5
3	5	2
4	3	4

Group_one	Group_Two	Group_Three
6	3	3
3	6	2
4	5	2

Group_one	Group_Two	Group_Three
4	4	4
4	4	2
2	7	3

Group_one	Group_Two	Group_Three
6	3	3
5	3	2
4	5	3

Group_one	Group_Two	Group_Three
4	6	2
5	3	2
5	3	3

Table 3.10. Performance of ML methods for predicting (T2-T0)/(T0), DB.v2

	PRECISION	RECALL	F1_SCORE	ACCURACY
Random Forest	0.35	0.45	0.38	0.45
SVM	0.29	0.51	0.37	0.51
K Nearest Neighbors	0.37	0.47	0.37	0.47
Decision Tree	0.42	0.41	0.41	0.41
Logestic Regression	0.38	0.45	0.40	0.45
MLP	0.38	0.44	0.40	0.44
Stochastic Gradient Descent	0.39	0.40	0.37	0.40
Adaptive Boosting	0.38	0.41	0.39	0.41
GradientBoosting	0.42	0.44	0.42	0.44
XGradientBoosting	0.41	0.44	0.41	0.44
Dummy Classifier	0.32	0.25	0.26	0.25

Table 3.11. Confusion matrix of multiclass classification, (T2-T0)/(T0), DB.v2

Group_one	Group_Two	Group_Three
16	0	0
7	0	0
6	1	0

Group_one	Group_Two	Group_Three
15	1	0
7	0	0
7	0	0

Group_one	Group_Two	Group_Three
13	0	3
6	0	1
6	0	0

Group_one	Group_Two	Group_Three
16	0	0
7	0	0
6	0	0

Group_one	Group_Two	Group_Three
15	0	0
7	0	0
7	0	0

In this step, we started to evaluate the performance of the ML methods in the case of binary classification. Table 3.12 shows the distribution of the target value in the case of binary classification. It should be mentioned although a 25% reduction is a cut-off for the (partial) response, we used 22% for having a more balanced database. Therefore, we used cut-off values to optimize the performance of the ML model and align with clinical criteria for patient response to treatment. This ensured accuracy and reliability, allowing the model to effectively address clinical objectives and produce consistent, desired outcomes.

Table 3.12. Distribution of (T2-T0)/(T0) for binary

>-0.23	<-0.22
85	84

For this experiment, the first version (DB.v1) of the database was used. Table 3.13 demonstrates that, in a small dataset like the current one, the ML methods performed better when binary classification was used. Moreover, according to the results, the performance of all classifiers was in the same range.

Table 3.13. Performance of ML methods for predicting (T2-T0)/(T0), DB.v1

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.47	0.47	0.47	0.47
SVM	0.45	0.45	0.44	0.45
K Nearest Neighbors	0.46	0.46	0.45	0.46
Decision Tree	0.49	0.49	0.48	0.48
Logistic Regression	0.44	0.44	0.44	0.44
MLP	0.45	0.46	0.45	0.45
Stochastic Gradient Descent	0.51	0.47	0.41	0.47
Ada Boost	0.48	0.49	0.48	0.49
Gradient Boosting	0.49	0.49	0.48	0.49
XGradient Boosting	0.45	0.45	0.44	0.45
Dummy Classifier	0.49	0.49	0.48	0.49

Table 3.14 presents the confusion matrix for the best ML algorithm. we visualized only the confusion matrices of gradient boosting since all of ML performances were similar.

Table 3.14. Confusion matrix of binary classification, (T2-T0)/(T0), DB.v1

Group_One	Group_Two
9	8
7	10

Group_One	Group_Two
9	8
8	9

Group_One	Group_Two
11	6
13	4

Group_One	Group_Two
8	9
10	7

Group_One	Group_Two
10	7
10	6

Table 3.15 compares the performance of various ML methods while they were predicting (T2-T0)/(T0) based on the DB.v2. The results indicated that adding additional features to the analyzing process had no significant influence on the outcome, and the performance was approximately the same.

Table 3.16 is showing the confusion matrix while predicting the target value using the decision tree algorithm. There is no doubt that this classifier is able

Table 3.15. Performance of ML methods for predicting (T2-T0)/(T0), DB.v2

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.49	0.50	0.49	0.49
SVM	0.49	0.48	0.46	0.48
K Nearest Neighbors	0.53	0.54	0.51	0.54
Decision Tree	0.54	0.54	0.54	0.54
Logistic Regression	0.41	0.44	0.41	0.44
MLP	0.49	0.49	0.48	0.49
Stochastic Gradient Descent	0.52	0.52	0.43	0.52
Ada Boost	0.43	0.43	0.43	0.43
Gradient Boosting	0.53	0.53	0.52	0.53
XGradient Boosting	0.48	0.48	0.47	0.48
Dummy Classifier	0.52	0.51	0.51	0.51

to correctly predict most cases by using the data related to the second version of the database (DB.v2).

Table 3.16. Confusion matrix of binary classification, (T2-T0)/(T0), DB.v2

Group_One	Group_Two
8	8
8	6

Group_One	Group_Two
6	10
7	7

Group_One	Group_Two
10	5
5	9

Group_One	Group_Two
11	4
7	7

Group_One	Group_Two
8	7
7	7

- **Predicting T2, Multiclass Classification**

As part of the second group of analyses, the value of T2 was assumed to be a label, and eleven different classifiers were used to predict it. The initial phase of forecasting the value of (T2) involved classifying it into three distinct categories (multi-class) based on HDRS score level (HDRS score level of depression: 10-13 mild; 14-17 mild to moderate; >17 moderate to severe), in this case instead 17 we used 22 as the cut-off between moderate to sever just to feed the algorithms with balance dataset and improve the performance ML methods. So, the dataset was as illustrated in Table 3.17 for feeding ML algorithms with a balanced dataset.

Table 3.17. Distribution of T2 for multiclass

<14	13< <23	>22
59	57	53

Table 3.18 compares the performance of eleven different ML algorithms while dealing with the small size of the dataset (DB.v1) and with the small number of predictors for predicting T2. This table demonstrates that adaptive boosting was capable of achieving better results than any other method. In comparison, the worst results were from the dummy classifier.

Table 3.18. Performance of the ML methods for predicting T2, DB.v1

	PRECISION	RECALL	F1_SCORE	ACCURACY
Random Forest	0.47	0.45	0.46	0.45
SVM	0.33	0.30	0.29	0.30
K Nearest Neighbors	0.45	0.38	0.37	0.38
Decision Tree	0.46	0.45	0.45	0.45
Logestic Regression	0.41	0.39	0.39	0.39
MLP	0.45	0.44	0.44	0.44
Stochastic Gradient Descent	0.47	0.41	0.38	0.41
Adaptive Boosting	0.48	0.48	0.48	0.48
GradientBoosting	0.47	0.45	0.46	0.45
XGradientBoosting	0.42	0.41	0.40	0.41
Dummy Classifier	0.32	0.31	0.31	0.31

Table 3.19 depicts the confusion matrix of adaptive boosting when it has been used to predict HDRS after ten treatment sessions. The table indicates that adaptive boosting can achieve a 17% higher performance than the dummy classifier.

The same kind of analysis for predicting T2 as (T2-T0)/(T0) had been done. Like the previous part of the analysis, this part also considered the code of the treating doctor and the various combinations of medication that were taken, as well as other factors (DB.v2). Afterward, all eleven ML methods were tested on how well they performed when they were being used to predict the Hamilton depression scale after tDCS. Table 3.20 shows the performance of all eleven ML

Table 3.19. Confusion matrix of multiclass classification, T2, DB.v1

Group_one	Group_Two	Group_Three
5	5	2
3	6	3
2	3	5

Group_one	Group_Two	Group_Three
6	2	3
2	5	5
3	4	4

Group_one	Group_Two	Group_Three
6	4	2
2	7	2
2	4	5

Group_one	Group_Two	Group_Three
4	5	3
6	4	1
3	1	7

Group_one	Group_Two	Group_Three
3	7	2
3	6	2
0	2	8

methods which were used in the analysis. According to the results obtained by the ML algorithms, with only a few percentage differences, all algorithms achieved similar accuracy ranging from 20 to 47 percent. The worst performance belonged to the dummy classifier, while the best classifier was gradient boosting.

Table 3.20. Performance of the ML methods for predicting T2, DB.v2

	PRECISION	RECALL	F1_SCORE	ACCURACY
Random Forest	0.41	0.41	0.41	0.41
SVM	0.32	0.31	0.31	0.31
K Nearest Neighbors	0.34	0.34	0.33	0.34
Decision Tree	0.41	0.41	0.41	0.41
Logestic Regression	0.38	0.37	0.37	0.37
MLP	0.40	0.38	0.37	0.38
Stochastic Gradient Descent	0.43	0.41	0.38	0.41
Adaptive Boosting	0.42	0.41	0.41	0.41
GradientBoosting	0.47	0.47	0.46	0.47
XGradientBoosting	0.42	0.43	0.42	0.43
Dummy Classifier	0.22	0.20	0.20	0.20

The obtained confusion matrix has been shown in Table 3.21 illustrates how gradient boosting achieved 47% of the performance.

- **Predicting T2, Binary Classification**

This section aims to apply binary classification, so the value of T2 was categorized into two groups. As 18 is the cut-off between moderate and severe depression

Table 3.21. Confusion matrix of multiclass classification, T2, DB.v2

Group_one	Group_Two	Group_Three
2	3	5
2	1	7
1	3	6

Group_one	Group_Two	Group_Three
5	5	1
2	7	1
1	4	4

Group_one	Group_Two	Group_Three
6	3	1
2	5	3
0	3	6

Group_one	Group_Two	Group_Three
5	4	1
6	2	2
1	1	7

Group_one	Group_Two	Group_Three
6	1	3
3	5	2
3	4	2

(HAM-D score level of depression: 10 - 13 mild; 14-17 mild to moderate; >17 moderate to severe), the values of T2 were divided based on it. Table 3.22 shows the distribution of T2 under binary classification.

Table 3.22. Binary distribution, T2

<18	>17
82	87

Table 3.23 shows the results obtained by all of the ML classifiers in the binary case. It shows that while the performance of all of those methods has improved, the accuracy of all applied ML methods was approximately similar. Their performance in terms of accuracy was in the range of 46 to 63 percent. While the best ML method was gradient boosting with 63%, the worst performance belonged to the support vector machine.

Table 3.24 illustrates the confusion matrix for all five-fold cross-validations belonging to gradient boosting, which was the best classifier compared to other methods. It could predict most cases correctly with the help of other information provided in the first version of the database.

The next step involved a combination of the medicine used and the doctors of the patients (DB.v2). The performance of the all-run ML method is depicted

Table 3.23. Results, binary classification, T2, DB.v1

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.60	0.59	0.59	0.59
SVM	0.42	0.43	0.42	0.42
K Nearest Neighbors	0.45	0.46	0.46	0.45
Decision Tree	0.58	0.58	0.58	0.58
Logistic Regression	0.57	0.56	0.56	0.56
MLP	0.56	0.55	0.54	0.55
Stochastic Gradient Descent	0.57	0.56	0.55	0.56
Ada Boost	0.59	0.59	0.58	0.59
Gradient Boosting	0.63	0.63	0.62	0.63
XGradient Boosting	0.60	0.59	0.59	0.59
Dummy Classifier	0.46	0.46	0.45	0.46

Table 3.24. Confusion matrix of binary classification, T2, DB.v1

Group_One	Group_Two
11	6
7	10

Group_One	Group_Two
8	9
4	13

Group_One	Group_Two
12	4
4	14

Group_One	Group_Two
8	8
9	9

Group_One	Group_Two
9	7
5	12

in Table 3.25. According to the results of the study, Extreme Gradient Boosting (XGBoost) obtained better performance compared to other ML methods. In

Table 3.25. Results, binary classification, T2, DB.v2

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.56	0.56	0.55	0.56
SVM	0.50	0.50	0.49	0.50
K Nearest Neighbors	0.47	0.48	0.46	0.48
Decision Tree	0.57	0.56	0.56	0.56
Logistic Regression	0.60	0.59	0.58	0.59
MLP	0.52	0.52	0.51	0.52
Stochastic Gradient Descent	0.54	0.54	0.50	0.54
Ada Boost	0.56	0.55	0.55	0.55
Gradient Boosting	0.58	0.56	0.55	0.57
XGradient Boosting	0.60	0.60	0.59	0.60
Dummy Classifier	0.52	0.52	0.52	0.52

Table 3.26, all confusion matrices related to the XGBoost were displayed. In this case, XGBoost was the best ML method for predicting target value by using the information of the second version database(DB.v2).

Table 3.26. Confusion matrix of binary classification, T2, DB.v2

Group_One	Group_Two
9	5
9	7

Group_One	Group_Two
8	7
7	8

Group_One	Group_Two
6	8
3	12

Group_One	Group_Two
8	6
5	10

Group_One	Group_Two
10	4
5	10

3.3.4 Discussion

Analyzing Hamilton Depression Scale scores post tDCS therapy can provide valuable insights for physicians in devising an effective treatment plan as part of the transition to Healthcare 4.0. Additionally, these kinds of clinical AI-based methods can help physicians offer more personalized counseling to their patients.

As part of this study, we collected demographic and clinical information about psychiatric patients who had been treated at the University of Naples, Federico II, both prior to and after receiving tDCS treatment. Due to this, we were able to examine whether the Hamilton depression scale score will be able to be predicted after treatment with tDCS. A preliminary phase of this study was the development and analysis of a methodology that could be used to process the input data and analyze the results in an objective way. To determine the variable at baseline that was associated with the Hamilton Depression Scale scores ten to twenty sessions after treatment, we evaluated the importance of each variable. Then, in order to determine the best machine learning classifier for this small dataset, we conducted several experiments.

The proposed models allowed us to identify the most predictive features (including clinical and demographic characteristics) that may assist in predicting the Hamilton Depression Scale and understanding the relationship between them. Further, based on the results of this study, it was found that the Hamilton depression scale score is potentially predictable if baseline data was taken into account. Aside from that, it has been shown that no single predictor can be used to achieve this goal. The current study identified that, among the variables collected during the baseline, the Hamilton depression scale score, along with the gender and age of the patients, were the best predictors of achieving the supposed goal. Additionally, when it came to handling small amounts of data, the results showed that all of the tested ML methods had similar performances, although up to 63% of accuracy was achieved by gradient boosting.

3.3.5 Conclusion

In order to evaluate the potential of AI for supporting clinicians in clinical decisions, we focused on two critical objectives. First, we propose the use of a ML approach to detect the most significant variables for predicting Hamilton depression scale levels after using tDCS therapy for treating depression. Moreover, the second goal of this work was to investigate the capability of ML classifiers to deal with the small size of the dataset. The analysis showed that the age, gender, and score of the Hamilton depression scale are the most significant predictors for predicting the level of the Hamilton depression scale after tDCS treatment. Moreover, the study showed that all methods have approximately the same level of performance, although gradient boosting achieved 63% of accuracy in one of the analyses.

However, future studies with a larger sample could be useful to confirm our findings. This model can be used to develop clinical AI-based systems such as clinical AI-DSS to plan and decide how to treat depressed individuals more efficiently.

Chapter 4

Metabolic Diseases

4.1 Introduction

In recent years, diabetics have become one of the modern challenges of the world. As a result of the interconnection between modern innovative technologies as well as advanced healthcare strategies, the concept of diabetes management has been redefined in the modern world. The purpose of this chapter is to provide a comprehensive summary of cutting-edge diabetes management in a different way. Specifically, it describes a dual-pronged approach involving both the influence of physical activity and machine learning-based predictive analytics to optimize glycemic control.

The first segment is dedicated to exploring the role of key enabling technologies that enhance the effectiveness of physical activity in managing diabetes. Specifically, we examined the role of wearable devices, smart monitoring systems, and physical activity in promoting diabetes management. Wearable devices, smart monitoring systems, and technological interventions have gained attention for improving health outcomes and reducing healthcare costs by providing real-time feedback on physical activity levels. These technologies can revolutionize how we approach physical activity and manage diabetes. Therefore, as part of our ongoing research in the field, we have undertaken a systematic literature review to comprehensively understand the role KET that has been studied by other researchers and to identify areas where further work is required to address gaps in this domain.

In this chapter also, the potential of using AI methods to predict glycemic events in diabetics is discussed. This can help to reduce the negative consequences of hypo or hyperglycemia events. The development and effectiveness of a prediction model are examined by incorporating various demographic information, clinical parameters, and analytical information. As a result, the paradigm of diabetes care can be redefined through early detection and preemptive interventions.

This scientific exploration highlights the potential of improving diabetes management by using new methods that employ AI, which aligns with healthcare 4.0.

4.2 Physical Activity Impact on Diabetes Management via Key Enabling Technologies: a Systematic Literature Review

4.2.1 Introduction

Diabetes Mellitus (DM) is a metabolic disease of inadequate control of blood glucose levels. DM is classified into two different ways Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D). The T1D refers to the fact that the body of people does not produce the required amount of insulin. This may end up receiving excess insulin, resulting in hypoglycemia (low blood sugar, which may result in brain damage, especially in the pediatric population) or low insulin production, resulting in hyperglycemia (high blood sugar may result in death). On the other hand, T2D means that the body cannot use insulin as well as it should, causing a high likelihood of hyperglycemia [138]. In any case, these conditions can lead to serious damage to the heart, blood vessels, eyes, kidneys, and nerves over time, which may lead to even death. According to the latest reports by the World Health Organization, in 2019, diabetes was the direct cause of 1.5 million deaths, and 48% of all deaths due to diabetes occurred before the age of 70 years [174]. However, effective management of diabetes can result in prolonged and improved quality of life. The effective management of diabetes also involves a balanced diet and regular physical activity as these lifestyle factors are interrelated and associated with insulin resistance and metabolic conditions [50]. The quantification and regular tracking of these lifestyle factors require modern KET such as portable wearables, smart watches, sensors, etc.

Studies using KET alone for diabetes management showed that physical activity improved the control of glucose levels [117, 53, 120]. However, with the advent of modern data-driven methods (such as machine learning approaches), their diffusion to KET is becoming widely explored in the field of diabetes for providing new and innovative ways to predict and manage glycaemic events and to estimate disease trajectory based on lifestyle factor tracking. In this section, we conducted a systematic literature review to examine the use of data-driven AI-based methods to improve diabetes management using data collected through

KET tracking physical activity and lifestyle quality for improved quality of life. This review aims to provide an understanding of the role that technology-based physical activity monitoring can play in managing diabetes.

4.2.2 Materials and Methods

The systematic literature review was conducted according to PRISMA guidelines [122]. This systematic literature review was carried out according to a PROSPERO protocol that defined and included the methods, and the criteria for inclusion and exclusion.

Search Strategy

Relevant titles were identified through searches in PubMed, Scopus, and Web of Science databases. we searched the publications from the last 10 years as there had been no significant involvement of KET for diabetes management and relatively older technology was utilized prior to 2012 which might not be in use today. we conducted the latest search in May 2023. we defined different combinations of keywords related to diabetes, artificial intelligence, KET utilized for tracking physical activity and other lifestyle factors. Table 4.1 shows how keywords from four subtopics were combined to find relevant papers. After defining research keywords related to the topic, we performed a comprehensive search in PubMed, Scopus, and Web of Science, searching for both the index test being evaluated and the target condition of interest. Two independent authors (SR, WHT) did an initial evaluation of all papers by analyzing their titles and abstracts against eligibility criteria following removing all repetitive papers by Python code, and all disagreements between authors were resolved by a third author (MSH) in a discussion.

Table 4.1. Search string(s) used to query databases

N	Keywords	Strings
#1	Diabetes and synonyms	(ALL (diabetes OR type 1 diabetes OR type 2 diabetes OR T1D OR T2D))
#2	Artificial intelligence and synonyms	(ALL (machine AND learning OR deep AND learning OR artificial AND intelligence OR clustering OR regression OR supervised AND classification OR unsupervised AND classification OR neural AND network OR reinforcement AND learning OR unsupervised AND learning OR supervised AND learning))
#3	Synonyms related to the KET	(ALL (sensor* OR smart AND watch* OR wearable* OR decision AND support AND system OR ecg OR ekg OR internet AND of AND thing OR iot OR medical AND devic* OR devic*))
#4	Physical activity and lifestyle synonyms	(ALL (physical AND activity OR exercise OR physical AND fitness OR workout OR fitness OR physical AND training))
#1 And #2 And #3 And #4		

Table 4.2. Inclusion criteria and exclusion criteria

Inclusion Criteria:
1. The papers which had used state-of-the-art methods based on artificial intelligence or regression for diabetes management 2. The aim of the suggested algorithms for data analysis in the study was managing diabetes 3. The paper used data from key enabling technologies as predictors in data analysis 4. Impact of physical activity on diabetes management
Exclusion Criteria:
1. Review articles, letters, abstracts, conference papers and case reports 2. The study only included healthy individuals 3. Training, learning, and/or validation process were not explained clearly or distinguished from each other 4. Non-human subjects (e.g., animals) 5. Non-English papers

Inclusion and Exclusion Criteria

Our key focus was to include journal articles using AI methods (e.g. regression models, machine learning, and deep learning methods) investigating the impact of physical activity on effective diabetes management utilizing KET. This led to the definition of inclusion criteria, which we enumerate as follows:

1. The studies must use the physical activity data acquisition utilizing KET (e.g. fitness trackers and smartphones).

2. Data acquisition of physical and electronic health records representing diabetic subjects.

3. Development and deployment of artificial intelligence techniques to analyze and model the data.

In the next step, we defined exclusion criteria to limit the search, such as:

1. Papers published before 2012 were excluded.

2. Studies involved other than diabetes, such as animals or healthy individuals, were excluded.

3. Studies published in other than English language were excluded.

4. Reviews, editorials, and articles were also excluded. However, we considered them to compare our findings for effective coverage of the scope of the topic.

we met the inclusion and exclusion criteria after mutual discussion among the authors and selected the research articles meeting the criteria. Table 4.2 details the criteria for inclusion and exclusion. For effective quality, two authors (SR and WHT) independently evaluated the eligible studies utilizing the aforementioned criteria. In case of disagreements, the eligibility of the studies was assessed by the third author (MSH).

Data Extraction

Following the completion of the evaluation by screening the title and abstract. A full-text review of selected literature was conducted to assess whether the papers met the inclusion criteria, at the same time, two authors extracted the relevant information (SR, WHT). The data extracted from the selected literature was as follows:

- Data collected from literature - The authors, publication date, study population, and data collection devices.
- Details of the data analysis - sample size, data analysis method, input variables, output variables, type of variable (continuous or discrete), the performance achieved.

Quality Assessment

Unlike statistical modeling methods, the quality of classification modeling methods are assessed in order to understand their robustness and trustworthiness when applied to the generalized population. After the data extraction, we performed a quality assessment of the classification modeling methods. The following tools were used when classification modeling methods were applied for data analysis:

- PROBAST (Prediction model Risk Of Bias Assessment Tool) - the PROBAST checklist tool was employed to assess the Risk of Bias (ROB) in the papers and to evaluate the applicability of diagnostic and prognostic predictions [173].
- TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) - The checklist TRIPOD was used to assess work with the aim of improving the reporting of studies involving the development, validation, or updating of predictive models for diagnostic or prognostic purposes [110].

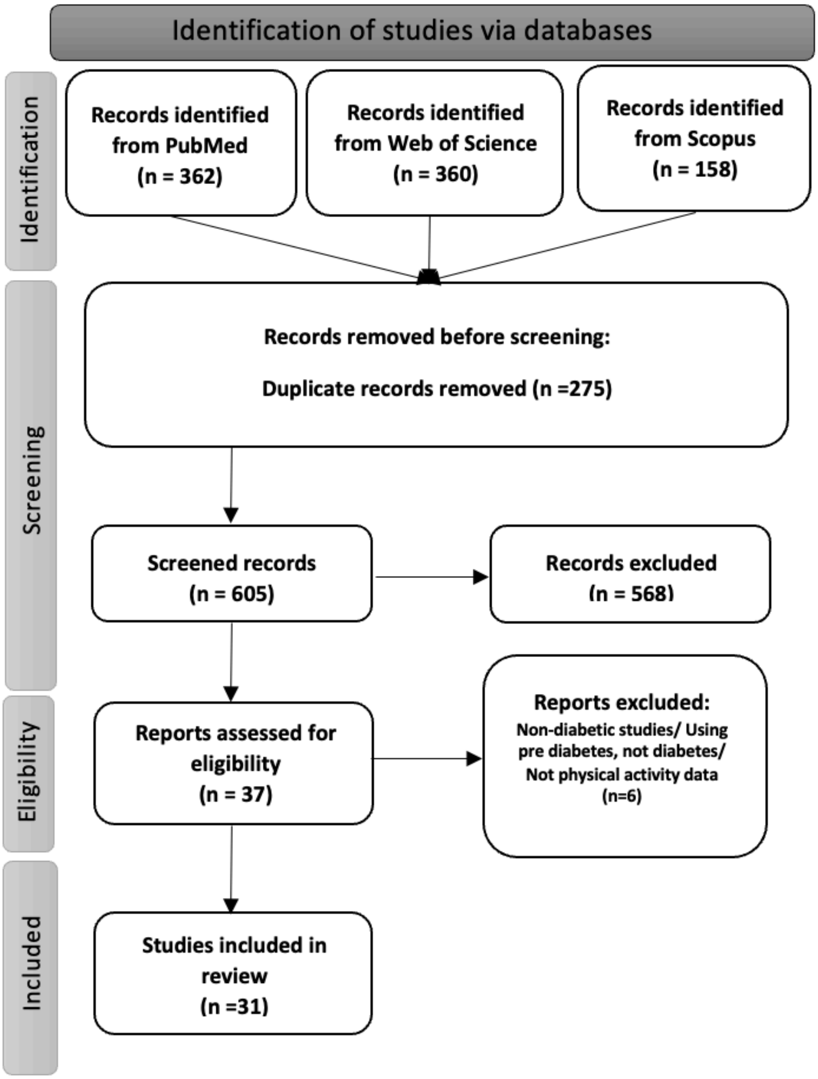

Figure 4.1. PRISMA search workflow

4.2.3 Results

In PubMed, Web of Science, and Scopus, 880 titles were identified using the above-described search strategy. Once duplications were removed, 605 papers were considered. The title and abstract of 568 papers mentioned in this paper were excluded because of the following reasons, 534 of the papers did not focus on physical activity and diabetes management directly, 27 papers were review articles, letters, abstracts, or conference papers, four papers worked only on healthy subjects, one paper studied animals, one was published in a non-English language, and one of them did not provide the clear explanation about training and validation process. In the next step, from 37 full-text articles remaining, six were excluded due to the exclusion criteria. The qualitative analysis and data extraction were conducted on 31 full texts. A flow chart of the literature search results is shown in Figure 4.1.

The articles selected for qualitative analysis and data extraction were divided into a) classification models and b) statistical methods. The statistical methods tend to determine associations between predictors and variables to be predicted. In this study of understanding physical activity impact on diabetes management, the predictors were physical activity and other lifestyle variables, whereas electronic health records (e.g. glycohemoglobin (HbA1c)) were variables to be predicted. On the other hand, classification methods aimed to predict diabetes and chronic event status based on predictors. The categorizing between statistical and classification methods was performed by SR and WHT whereas any disagreements were resolved by MSH.

The distribution of these papers is presented in Figure 4.2. This figure shows that approximately 35% of the papers used classification methods to diagnose different manifestations related to diabetics while they were doing physical activity and using KET to collect data. For data analyzing and predicting or diagnosing, they developed and deployed ML and deep learning methods like support vector machine, decision tree, random forest, eXtreme Gradient Boosting, or artificial neural networks. In the second group of papers which used statistical methods, the main goal was to find the association between different variables related to diabetes when they were doing physical activity and using key enabling technologies for gathering data. Research studies focusing on statistical methods for finding correlations accounted for about 39% of the studies. On the other hand, about 26% of the papers used both methods like statistical to find a correlation between variables as well as classification methods to predict or diagnose diabetics as they engaged in physical activity. The details of these methods are presented in subsequent subsections.

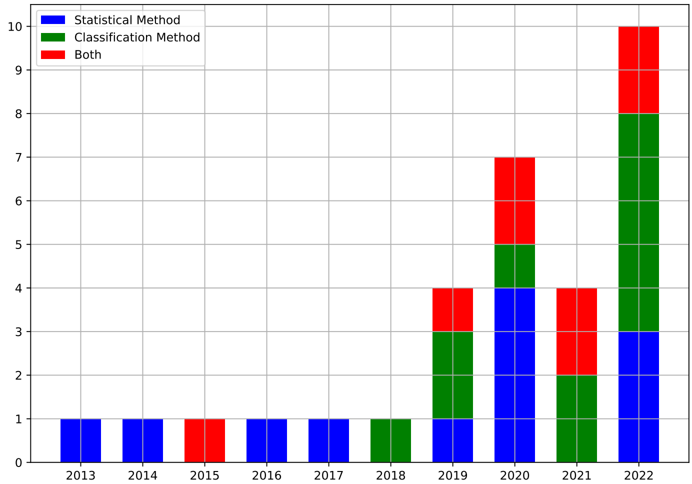

Figure 4.2. Number of publications by year

Characteristics of the studies

Figure 4.2 illustrates the popularity of analyzing diabetes and the impact of physical activity acquired through KET in order to manage the disease by year. It shows that the tendency among researchers to work in this field increased during the last few years.

In Table 4.3, the devices, data acquisition methods, the samples, whether the work included both healthy and diabetic subjects or just diabetic, the kind of variables used, whether discrete, continuous, or a mix of discrete and continuous variables and the hospitals where the studies have been conducted are outlined. Data represented by discrete variables is countable, whereas data represented by continuous variables is uncountable and infinite. It should be mentioned that data from the work led by Georga et al., [55] was driven by the study which was led by Zarkogianni [175]. The information in the table proves that the common devices are wearable devices like smartwatches. For example, the main smart device for collecting data by the research group led by Turksoy et al., was the SenseWear physical activity (PA) monitor [164], a wrist-worn device for monitoring the physical activity of the participants. Also, in another study led by Zarkogianni et al., the researchers used a SenseWear armband to collect data on the physical activity of the participants in the study [175]. In another study, Fitbit Inspire Heart Rate (HR) was used as the smart device used for collecting data from the participants. Christian et al. used both continuous and discrete types of data within their dataset in order to gain a better understanding of the role of physical activity in diabetes management [27]. As another type of smart device, ActiHear was also used by the research group led by Andrew et al. to gather data for a study that included both continuous and discrete variables [36]. It also shows that most of the data comes from the United States, even though some of the works did not mention the country where the data were collected. In the same field of research, the researchers from Oregon Health & Science University were trying to understand how physical activity and diabetes can be associated with each other based on the data gathered from type 1 diabetes mellitus participants who were monitored with a continuous glucose monitor smart device [131]. Another study conducted at Michael Kahn Washington State University (US) used smartphones to collect data from the participants, utilizing both continuous and discrete variables [9]. According to Table 4.3, the majority of the researchers preferred to use continuous variables rather than just discrete variables in their study. Accordingly, the table proves that there are two groups that use continuous variables or both continuous and discrete variables. A study conducted by Taiyu et al. in the research they conducted at Imperial College Healthcare NHS Trust used only continuous variables to investigate the impact of physical activity on the management of diabetes. In their work, they used Empatica E4 as a smart device for collecting data [179]. On the other side, Denes et al. used both continuous and discrete data as the variables. They used Zephyr BioHarness 3 and Continuous Glucose Monitoring (CGM) for collecting data and investigating [40].

Studies using regression and statistical models

Table 4.4 shows which kind of statistical method has been employed for analyzing the data for each research study. Linear regression was the most commonly used method to analyze datasets to determine whether variables were associated. Mixed effect linear regression and multivariable linear regression can be described as the second most common method. It should be noted that although some of the papers specified their regression model with all details, there were papers that only mentioned linear regression models without specifying what type of linear regression they used; thus, they were counted as just linear regression.

In the paper led by Muntis et al., the authors used mixed effects regression models to investigate the association of self-reported total weekly minutes of moderate-to-vigorous physical activity on each of the primary outcomes like HbA1c. In another paper, physical activity data was collected from T2D subjects to investigate the relationship between medical condition, medication, smoking habits, presence of diabetic polyneuropathy, muscle mass, protein intake, blood

Table 4.3. Characteristics of the reviewed paper. "C" refers to continuous variables, "D" to discrete variables, and "Both" means the study used both types of continuous and discrete variables.

Study Authors	*Device*	*Input Variable*	*Hospital*	*Samples*
Franklin et al. [117]	Garmin Vivosmart® 4	Both	Not specified	T1DM: 38
Jordan et al. [53]	Abbott Libre Pro, Metria IH1	C	Ten US diabetes centers	T1DM: 52
Christian et al. [27]	Fitbit Inspire HR	Both	Melbourne, Australia	T2DM: 11
Moser et al. [113]	rtCGM	C	Not specified	T1DM: 7
Ozaslan et al. [120]	Unspecified CGM, Fitbit Charge, HR and HR2	Both	Not specified	T1DM: 37
Hagedoorn et al. [67]	Fitbit Flex	Both	Ziekenhuisgroep Twente (Netherlands)	T2DM: 217
Henson et al. [77]	GENEActiv	Both	Midlands, UK	T2DM: 635
Laguna et al. [94]	Fitbit Charge HR, Microsoft Band 2, Polar HR monitor, Veo insulin pumps, Enlite-2 CGM	C	Clinic University Hospital of Barcelona (Spain)	T1DM: 6
Turksoy et al. [164]	SenseWear PA monitor	C	University of Illinois College of Nursing (US)	T1DM: 26
Hajna et al. [68]	GPS monitor, ADXL345	Both	Montreal (Canada)	T2DM: 97
Katrin et al. [42]	Qiu	C	Neurological group practice (Germany)	T2DM: 17
Andrew et al. [36]	ActiHeart	Both	Not specified	T2DM: 394
Sevil et al. [145]	Empatica E4	C	Not specified	Total: 25
Czmil et al. [38]	ActiGraph wGT3X-BT	Both	Rzeszow State Hospital (Poland)	Total: 330 T1DM: 215
Pieter et al. [78]	Shimmer®3inertial moment sensors	C	Tallaght University Hospital (Ireland)	Total: 138 T2DM: 94
Taiyu et al. [179]	Empatica E4	C	Imperial College Healthcare NHS Trust (UK)	T1DM: 12
Kyoung et al. [89]	Fitbit	C	Korea University Anam Hospital (Korea)	T2DM: 24
William et al. [166]	iPro2, Enlite, activPAL	C	Not specified	Total: 851 T2DM: 197
Lam et al. [95]	Wrist-worn triaxial accelerometers	Both	Not specified	Total: 10,129 T2DM: 1,666
Elhadd et al. [46]	Fibit-2, Libre	Both	Hamad Medical Corporation (Qatar)	T2DM: 13
Bertachi et al. [25]	FSL sensor, Fitbit Alta HR	Both	Not specified	T1DM: 10
Reddy et al. [131]	Unspecified CGM	Both	Oregon Health & Science University (US)	T1DM: 55
Alfian et al. [9]	Smartphone	Both	Michael Kahn, Washington University (US)	Total: 768 T2DM: 268
Zarkogianni et al. [175]	Guardian® Real-Time CGM, SenseWear Armband	C	Not specified	T1DM: 10
Cescon et al. [30]	Three-axis accelerometer	Both	Sansum Diabetes Research Institute (US)	T1DM: 20
Georga et al. [55]	Guardian® Real-Time CGM, SenseWear® Armband	C	Parma University Hospital	T1DM: 15
Sevil et al. [144]	Dexcom G5 CGM, E4 wristband	C	University Illinois (US)	T1DM: 12
Srinivasu et al. [156]	Gold oxide sensor	C	Not specified	600 spectrogram images
Rashid et al. [129]	Unspecified CGM	C	Not specified	T1DM: 50
Askari et al. [12]	Empatica E4, Bioplux, COSMED K5	C	Not specified	Total: 34 T1DM: 28
Dénes et al. [40]	Zephyr BioHarness 3 and Unspecified CGM	Both	Not specified	T1DM: 20

pressure, and physical activity [77]. Laguna et al., in their work, used linear regression for assessing the association between HR and the signal named FHR stands for F from the Fitbit Device, such as number of floors of stairs climbed, Fitbit Metabolic Equivalent of Tasks (METs), calories burned, number of steps, accumulated movement magnitude, polar heart rate (PHR); PHR measures heart rate using two different methods, optical heart rate measurement, and a chest strap heart rate sensor. To collect data, they used a range of KET devices [94]. Turksoy et al. carried out a study with the aim of examining the association between heart rate, heat flux, skin temperature, near-body temperature, galvanic skin response, energy expenditure, and 2D acceleration of absolute difference. To reach their goal, they used Partial Least Squares linear regression [164].

Table 4.4. Different regression methods used. "LR" refers to Linear regression, and "LRM" to the logistic regression model.

Author	*LR*	*Mixed-effects LR*	*Multivariate LR*	*Least Squares LR*	*LRM*
Franklin et al. [117]		X			
Jordan et al. [53]	X				
Christian et al. [27]				X	
Moser et al. [113]	X				
Ozaslan et al. [120]		X			
Hagedoorn et al. [67]			X		
Henson et al. [77]			X		
Laguna et al. [94]	X				
Turksoy et al. [164]				X	
Hajna et al. [68]	X				
Katrin et al. [42]	X				
Andrew et al. [36]			X		
Sevil et al. [145]	X				
Czmil et al. [38]	X				X
Pieter et al. [78]			X		
Lam et al. [95]					X
Elhadd et al. [46]	X				
Zarkogianni et al. [175]	X				
Cescon et al. [30]					X
Dénes et al. [40]					X

Studies using classification methods

The second group of papers utilized classification methods to examine their dataset to predict the impact of physical activity on diabetes management. Table 4.5 shows the common classification methods used for analyzing datasets. As shown, most of the papers used several methods, analyzing their performance for the classification of diabetes management phenomena. A glance at the table proves that Artificial Neural Network (ANN) methods are the most common way among researchers for analyzing datasets. The second common method for analyzing the dataset was using the SVM and Long Short-Term Memory (LSTM).

In Sevil et al. [145] work they used several methods like kNN, SVM, DT, Naive Bayes (NB), ANN, LR, and Deep Neural Network (DNN) model with long-short term memory to predict energy expenditures of the physical states. Another study that has used DNN was the work led by Elhadd et al.[46]; although they did not specify the details related to their DNN, they applied other models such as RF, SVM, and XGBoostto their dataset as well. Another paper developed the Recurrent Neural Network model method for predicting the glucose level of diabetics using sequential data from Fitbit, continuous glucose data, and subject demographics (BMI, Age, gender, etc.) [179]. Furthermore, Lam et al. used RF, LR, and XGBoost algorithm methods to predict whether participants had T2D. In their study, the data collected by wrist-worn triaxial accelerometers were used as predictors, and their models were accurate between 70% and 86%. Zarkogianni et al. [175] applied feedforward neural network (FNN), a self-organizing map, a neuro-fuzzy network with wavelets as activation functions, and a linear regression model for predicting the blood glucose of participants. For collecting data the authors used SenseWear Armband (BodyMedia Inc.) as key enabling technologies. Taiyu et al. used the method ANN method for their dataset [179].

Table 4.6 contains an overview of the main inputs and outputs of the AI model as well as the related performance. Furthermore, the table shows that most of the papers developed a model for predicting glucose levels based on the predictors proposed by the authors. The study that used interstitial glucose concentration, meals, insulin doses, self-monitoring blood glucose measurements (SMBG) values, HR, steps performed, estimation of calories burned, and sleeping period as the predictors to predict nocturnal hypoglycemia had the highest and lowest accuracy among all the research works whose accuracy has been specified, respectively, of 100 and 62 percent [25]. In the paper led by Cescon et al. showed that the models were accurate to the tune of 99.99%, which can be considered one of the best available performances [30].

Evaluating quality of the papers

This section evaluates the quality of papers that employed classification methods by utilizing PROBAST and TRIPOD, which were previously discussed.

We explain the quality assessment results utilizing both of these techniques as follows:

- **PROBAST Tool**

A Risk Of Bias Assessment Tool was applied to evaluate diagnostic and prognostic prediction models in order to determine their validity and reliability [173].

Table 4.5. Common classification methods used

Author	*kNN*	*DT*	*RF*	*NB*	*SVM*	*LSTM*	*XGBoost*	*ANN*	*DNN*
Sevil et al. [145]	X	X		X	X	X		X	X
Czmil et al. [38]		X	X		X			X	
Pieter et al. [78]								X	
Taiyu et al. [179]						X			
Kyoung et al. [89]									
William et al. [166]					X	X	X	X	
Lam et al. [95]			X				X		
Elhadd et al. [46]			X		X		X		X
Bertachi et al. [25]					X			X	
Reddy et al. [131]		X	X						
Alfian et al. [9]						X		X	
ZarkogiANNi et al. [175]								X	
Cescon et al. [30]	X		X	X	X	X		X	
Srinivasu et al. [156]								X	
Rashid et al. [129]						X		X	
Askari et al. [12]						X		X	
Dénes et al. [40]	X	X	X	X	X			X	

In addition, it facilitates the possibility of generalizing study findings to real-life settings by assessing the applicability of the results to the intended population. Study design, patient selection, outcome measurement, and statistical analysis are all aspects of study quality that were evaluated with this tool. Participant, predictor, outcome, and analysis are the four domains of PROBAST. Table 4.7 contains a list of questions related to each of the domains of PROBAST.

The tool is composed of 20 questions to evaluate the quality of the studies. Moreover, in order to ensure the reliability and trustworthiness of the results, these signaling questions were used as a structured approach to assess the potential risk of bias. Utilizing these questions, the methodology and implementation aspects of the study, such as choosing participants, measuring and analyzing outcomes, handling missing data, and analyzing results, have been evaluated. The questions corresponding to these domains were systematically examined.

Figure 4.3 presents the answer to the questions used to assess bias risks for each state-of-the-art. The figure also provides a detailed overview of the distribution of responses to each question, providing a comprehensive view of the assessment process. Study quality can be assessed using this information, and areas that could benefit from further investigation can be identified. As illustrated in Figure 4.3, most of the responses have been answered "yes" or "probably yes", indicating the authors' experiments were sufficiently detailed and that their chances of bias were low. From a glance at the participants, it appears that almost all papers utilized appropriate inclusion and exclusion criteria and

Table 4.6. Classifications and outcome of the papers used prediction method

Author	*Predictors*	*Outcomes*	*Accuracy*
Sevil et al. [145]	Skin temperature, 3-dimensional accelerometer, galvanic skin response, blood volume pulse, heart rate	Energy expenditures of the physical states	75.7%-94.8%
Czmil et al. [38]	Step count, sedentary, light, moderate, and vigorous activity minutes	Sick or healthy	86.9%-80.8%
Pieter et al. [78]	Gait variables: heel strike, toe-off, max swing velocity	Cognitive, neuropsychological assessment	Not specified
Taiyu et al. [179]	Electrodermal activity, inter-beat intervals, acceleration, skin temperature, daily meal composition, insulin injection, exercise, health conditions	Glucose level	88.6%-87.2%
Kyoung et al. [89]	Physical activity, heart rate, sleep	Glucose level	Not specified
William et al. [166]	Physical activity, Glucose level	Glucose level	Not specified
Lam et al. [95]	Number of bouts, time spent in the activity, length of the bouts	T2DM or not	70%-86%
Elhadd et al. [46]	Fitbit data, continuous glucose	Glucose level	82.6%-95.2%
Bertachi et al. [25]	Interstitial glucose concentration, meals, insulin doses, SMBG values, heart rate, steps performed, calories burned, Sleeping duration	Nocturnal hypoglycemia	MLP: 63%-99% SVM: 62%-100%
Reddy et al. [131]	Exercise heart rate, Energy expenditures, Glucose level, insulin at start of exercise, daily insulin dose, glucagon	hypoglycemia	DT: 78% RF: 97%
Alfian et al. [9]	Insulin dose, Glucose level, meal ingestion, physical activity	Glucose level	77%
Zarkogianni et al. [175]	Recent glucose level, glucose level fluctuation	Glucose level	Not specified
Cescon et al. [30]	Fitness condition, type and intensity sedentary, household, lifestyle, sport and gym activities	behavior classification	99.99%
Georga et al. [55]	Signal data	Glucose level	70%-99%
Sevil et al. [144]	Physical activity, acute psychological stress , meals, insulin	Glucose level	Not specified
Srinivasu et al. [156]	Spectrogram images	Glucose level	75%-81%
Rashid et al. [129]	CGM and insulin pump data	behavior classification	94.69%
Askari et al. [12]	3D accelerometer signal, BVP signal (body temperature, skin conductivity, and cardio vascular activity)	behavior classification	96.82% - 99.99%
Dénes et al. [40]	BG level from CGM (concentration), HR value (integer), Exercise (0/1)	behavior classification	29.9% - 84.5%

Table 4.7. PROBAST Questionnaire

Domain	**ID**	**Question**
Participants	Q1	Were appropriate data sources used?
	Q2	Were all inclusions and exclusions of participants appropriate?
Predictors	Q3	Were predictors defined and assessed in a similar way for all participants?
	Q4	Were predictor assessments made without knowledge of outcome data?
	Q5	Are all predictors available at the time the model is intended to be used?
Outcome	Q6	Was the outcome determined appropriately?
	Q7	Was a prespecified or standard outcome definition used?
	Q8	Were predictors excluded from the outcome definition?
	Q9	Was the outcome defined and determined in a similar way for all participants?
	Q10	Was the outcome determined without knowledge of predictor information?
	Q11	Was the time interval between predictor assessment and outcome determination appropriate?
Analysis	Q12	Were there a reasonable number of participants with the outcome?
	Q13	Were continuous and categorical predictors handled appropriately?
	Q14	Were all enrolled participants included in the analysis?
	Q15	Were participants with missing data handled appropriately?
	Q16	Was selection of predictors based on univariable analysis avoided?
	Q17	Were complexities in the data accounted for appropriately?
	Q18	Were relevant model performance measures evaluated appropriately?
	Q19	Were model overfitting, underfitting, and optimism in model performance accounted for?
	Q20	Do predictors and their assigned weights in the final model correspond to the results from the reported multivariable analysis?

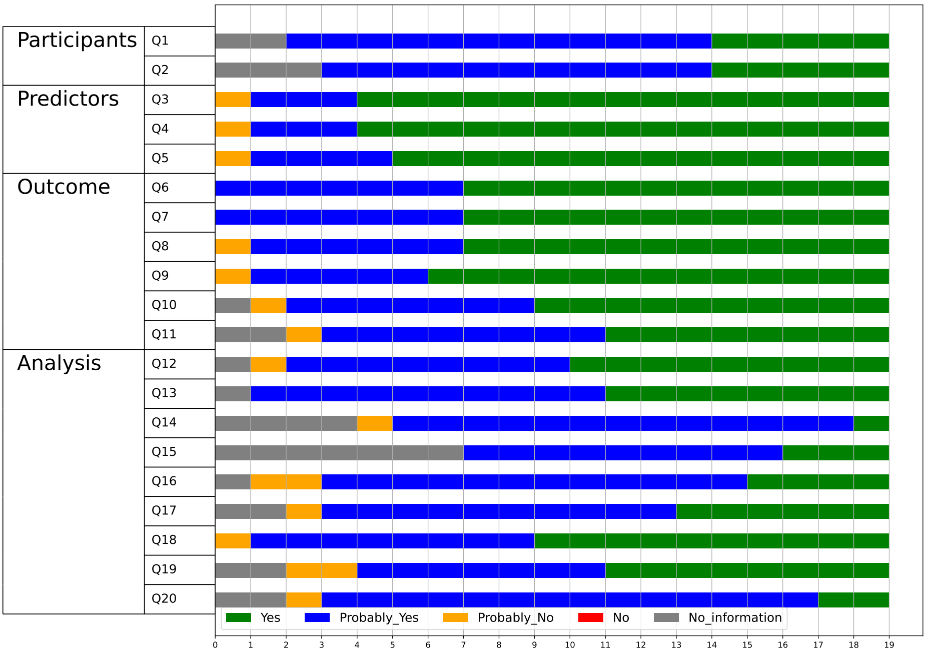

Figure 4.3. The proportion of Risk of Bias

data sources. It is important to note that most papers defined "predictors" in the same way for all participants, and they were also assessed without knowledge of the outcome and were available at the time of use. Moreover, Figure 4.3 illustrates that defined outcomes were used for evaluating the AI methods, they were defined properly, and all the participants were assessed identically. The last part of the figure shows the questions related to the analysis method. A preliminary look indicates that approximately half of the papers do not mention how they dealt with missing data. It has been observed that most authors provide a detailed explanation of their method of analysis in their papers, indicating that the risk of bias is low. Figure 4.4 and 4.5 are focused on the risk of bias and applicability of studies, respectively. these figures provide explanations and elaborations for PROBAST, which can be used to assess the risk of bias and applicability of studies developing, validating, or updating prediction models for individualized diagnostics or prognoses.

Referring to Figure 4.4 proves how many papers are classified as high risk of bias or low risk. It shows that studies in this area had an adequate number of participants and low bias risks, according to the results. However, the analysis revealed that only one or two papers had a high risk of bias, depending on the domain of the Probast analysis, while approximately 90 percent of the papers had a low risk of bias. According to the last row of the figure, most authors minimized bias by taking appropriate measures, whereas one paper did not provide adequate information on one or more of the concerned domains. Further, one paper was recognized as having a high risk of bias because there were one or more areas that were deemed to be at high risk of bias.

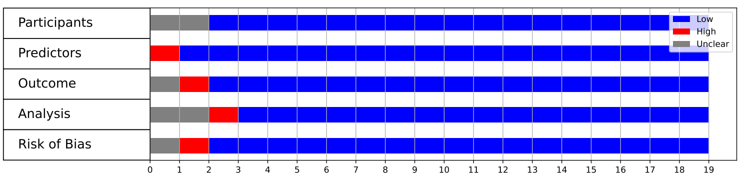

Figure 4.4. The proportion of papers in terms of risk of Bias

Aside from assessing the study risk of bias, PROBAST determines whether the results were applicable to the target population. If the population, predictors, or outcomes of a primary study differ from those specified in the review question, there may be concerns regarding the applicability of the study to the review question [173]. A prediction model study involves assessing the following four domains: participants, predictors, outcomes, and analyses. According to Figure 4.5, all the papers were found to be applicable in terms of participant inclusion criteria. As compared with the other three domains, a lower proportion of papers were considered applicable when evaluating their outcomes. Nonetheless, the majority of the reviewed papers were applicable, with more than 80 percent meeting the applicability criteria.

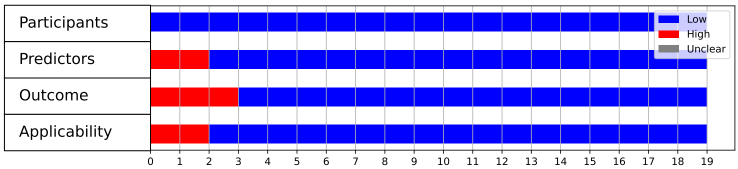

Figure 4.5. The proportion of papers in terms of applicability

- **TRIPOD Tool**

According to the TRIPOD statement, studies using prediction models can be divided into five categories, including: 1. prognostic or diagnostic predictor finding studies, 2. prediction model development studies without external validation, 3. prediction model development studies with external validation, 4. prediction model validation studies, and 5. model impact studies. For conducting TRIPOD the first step consists of finding prognostic or diagnostic predictors, followed by developing prediction models without external validation, and finally, evaluating the impact of the models.

According to the TRIPOD Statement, studies developing or validating multivariable prediction models must follow a checklist of 22 items thought to be essential. The items of TRIPOD, which cover key aspects of study design, are related to the title and abstract, background and objectives, methods, results, discussion, and other information. Whether for diagnostic or prognostic purposes, TRIPOD aims to improve the reporting of studies that develop, validate, or update prediction models [110]. Moreover, according to the TRIPOD Statement, despite the method of study, the transparency of reporting a prediction model study should be improved.

Two questions regarding the study title and abstract were included in the TRIPOD checklist initial domain. According to Figure 4.6, roughly 80 percent of the responses indicate that the title and abstract adequately reflect the content of the study, with a "Yes" response. The second domain of TRIPOD focuses on the introduction section, which consists of two questions that assess whether the medical context and main objective of the study are appropriate. As shown in Fig 4.6, most studies (over 88%) provided adequate details about these aspects.

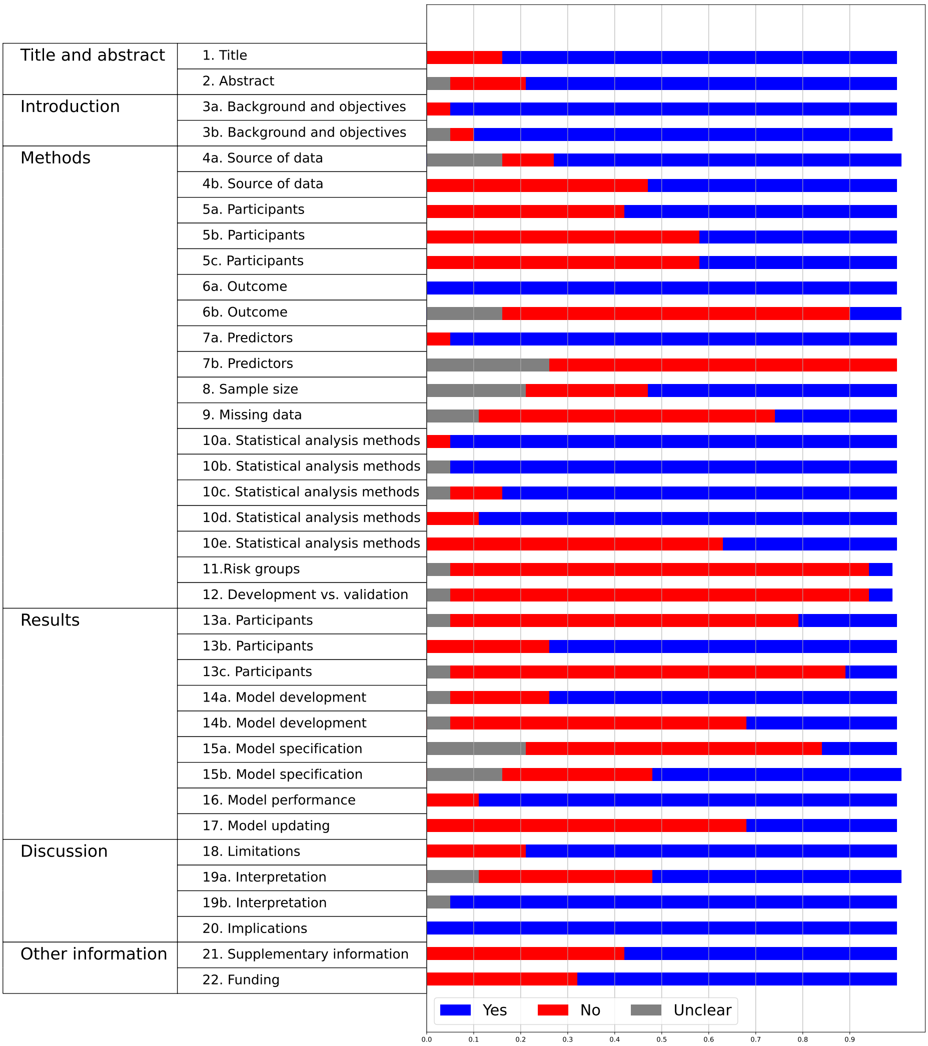

Figure 4.6. The proportion of papers in terms of TRIPOD assessment

A domain of the TRIPOD checklist evaluates research methodology sections of papers. This domain evaluates the extent to which the study explains the source of data, participants, outcome, predictors, sample size, missing data, statistical analysis methods, risk group, and whether it was developed or validated. As shown in Figure 4.6, most studies did not adequately address some areas. Based on the results of question number 7b, most studies did not mention the blinding of predictor assessments for outcomes, or the information was unclear when it was mentioned. This indicates why most of the answers to question 7b were negative. Meanwhile, according to question number 6b, all papers analyzing their dataset using predictive methods clearly define the outcome that is predicted by the prediction model, as well as how and when it is evaluated. According to questions 11 and 12, only seven percent of the papers provided information about managing the risk group, approximately 86 percent of the works did not provide any information, or the explanation related to the risk group was unclear for the remaining papers (7%). Another domain of TRIPOD is related to the result. According to the Figure 4.6 result section, only five percent or less of the studies provided unclear information about their subjects. Furthermore, more than 85 percent of the papers clearly reported performance measures based on answers to question number 16. The next domain, which was covered by the TRIPOD tool, was the discussion section of the papers. In general, the discussion part of Figure 4.6 proves that over 58% of the papers met requirement items related to this section of TRIPOD. The result related to question number 20 illustrates that almost all papers discuss the potential clinical use of the model and its implications for future research. The papers also clearly stated their research limitations in 79 percent of them. The last domain covered by TRIPOD is titled "Other information". This section of the paper investigates whether they disclose their source of funding and supplementary resources, and the figure shows that over half of the papers covered it.

4.2.4 Discussion

Diabetes management and its relation with physical activity have become one of the pressing issues of recent years. Therefore, studies were conducted in order to find a solution for understanding it. The result of this work showed that the most commonly used devices for data collection were wearable devices such as different Fitbit types, although there are also some works that use other devices such as CGM. Based on the current investigation, most of the papers did not mention the challenges associated with collecting data through KET, some of them mentioned that due to the misuse of KET or failure to adhere to research rules, some data had to be discarded. Regarding the type of variable used, it was found that most of the papers worked with continuous variables and only a few with a mixture of discrete and continuous variables; on the other hand, none of the papers worked only with discrete variables. By analyzing the papers focused on this area, we identified two groups of work among those studies. The first group used statistical methods to analyze their dataset and find the relationship between different variables related to diabetes management while engaging in any kind of physical activity. Here, the authors mainly evaluated the relationship between different physical activity variables and blood glucose levels. Most of the studies performed in 2020 and later showed high popularity of measuring the impact of physical activity with key enabling technology on diabetes trajectory. The results also demonstrated multivariate linear regression as one of the most common methods for analyzing datasets among researchers to reach their research goal and find the correlation between variables.

The second group used predictive methods to analyze data to predict or diagnose diabetic-related events. During the last ten years, there have been a number of papers that have analyzed the dataset using predictive methods, with the first published in 2015 and most published after 2020. SVM and RF are the most popular among AI methods for analyzing datasets. Moreover, the researchers aimed at predicting blood glucose levels as their main objective, followed by predicting whether participants had diabetes. A comparison of the results of all the AI methods used revealed that ANN and LSTM achieved the best results in terms of classifying behavior. A further finding of the current study is that physical activity and blood glucose levels are the most common predictors for determining the impact of physical activity on diabetes.

To determine whether the predictive method would be bias-free and applicable, we evaluated the papers that used the predictive method using the PROBAST tool. Generally, the papers met most of the critical requirements for having a low bias risk and applicability. It was discovered that more than half of the papers fulfilled most of the required criteria, despite some weaknesses in covering the method and result section, where only a few papers covered the details clearly. Some papers had unclear information about the analysis, outcome, and participant domains. Most of the domains covered were adequately detailed and were considered to have a low bias risk, according to PROBAST. Results of assessing papers with the PROBAST checklist indicate that most of the papers match the guidelines that this checklist developed for evaluating papers in terms of applicability and bias.

TRIPOD was another tool to evaluate papers that used predictive methods. All five domains of the TRIPOD checklist were used to evaluate this group of papers, including title, abstract, background and objectives, methods, results, discussion, and other information. While more than 50 percent of the work in this area meets the majority of the needs of TRIPOD, there are some domains, such as method and results, in which more attention needs to be given, and more details need to be provided by the researchers.

4.2.5 Conclusion

As a result of the human lifestyle, diabetes has become one of the most important issues over the past few years. It is important that researchers who work in this field pay more attention to physical activity and its effect on diabetes management as one of the main treatments for diabetes. Therefore, a systematic literature review of the papers investigated the detection of the impact of physical activity on diabetic management via key enabling technologies. Research on this topic has become increasingly popular in the past four years.

In the current study, it was found that while diabetics are performing any physical activity, some features, such as glucose level, can be predicted based on data obtained from KET and analyzed by AI methods. The papers that were investigated in this study demonstrated an overall correlation between the management of diabetes and physical activity, as physical activity can be used as an effective, non-invasive method to control the blood glucose level in the body.

we can conclude from this study that two methods were used to analyze the impact of physical activity on diabetes management: one based on statistical methods and one based on classification methods. In the PROBAST tool evaluation of papers based on the classification method, most papers and areas were found to have a low risk of bias. According to the TRIPOD checklist assessment results, although some domains need more attention, most have been adequately covered by the authors or the works.

Current research indicates that, although there are papers addressing the impact of physical activity on diabetics, This area requires more attention from researchers as a key issue in modern life. Moreover, researchers should reduce the risk of bias by being more precise. Additionally, the TRIPOD criteria should be followed more closely in some domains, particularly in the method and result areas.

Given the prevalence and impact of diabetes in modern society, it is important to develop AI-powered systems that can improve the quality of life for individuals with diabetes, enabling them to live a lifestyle similar to those without the condition. In this regard, research should be conducted in the field of Healthcare 4.0 that can be used to personalize the methods of diabetes management according to the level of physical activity of the individual; and it should be able to predict the blood glucose level based on the individual characteristics profile and physical activity.

4.3 Prediction of Glycemic Event in Emergency Section Patients Using Machine Learning

4.3.1 Introduction

The Centers for Disease Control and Prevention classifies chronic diseases as long-term conditions that persist for a year or more, require ongoing medical care, and can limit daily activities [2]. Cardiovascular disease, cancer, chronic respiratory conditions, and diabetes are the most prevalent chronic diseases. They not only cause disability and deterioration of health but also contribute to an increase in premature mortality in the European Union (EU). According to the Organisation for Economic Co-operation and Developmen (OECD), about 550,000 people of working age in the EU succumb to these diseases every year [3]. As the leading cause of mortality, chronic diseases are responsible for the largest share of healthcare expenditure, with an estimated cost to EU economies of €115 billion per year, or 0.8% of Gross Domestic Product (GDP) [35].

In this respect, DM represents a major challenge for the world's healthcare systems, as its prevalence and associated healthcare costs are steadily increasing. Diabetes, which is not limited to older adults, is a major cause of healthcare expenditure and a significant determinant of life expectancy in Europe, with a current prevalence of 2.8% across all age groups worldwide; this figure is expected to roughly double by 2030 [159, 137]. DM is a chronic metabolic disorder characterized by insufficient insulin production Type 1 Diabetes Mellitus (T1DM) or impaired insulin sensitivity Type 2 Diabetes Mellitus (T2DM), leading to prolonged periods of high or low blood glucose levels [178].

Patients with diabetes may experience hypoglycemia or hyperglycemia, which can cause complications in vital organs [154, 138].

The conjunction of digitization and the integration of artificial intelligence AI into healthcare has meant that historical data and AI-based algorithms can now discriminate complex patterns and predict episodes of hyperglycemia and hypoglycemia, which enables personalized diabetes management and optimizes patient outcomes.

According to previous research, it is possible to predict future glucose levels by following a certain pattern of recent glucose measurements [154]. Moreover, it has been found that severe hypoglycemia often follows a specific blood glucose fluctuation pattern that is recognizable from routine self-monitored blood glucose levels [37]. Fraser Cameron et al. have shown that statistical linear prediction provides significant lead time in predicting hypoglycemic events prior to their occurrence [29]. In another research study, the authors used real-time data from continuous glucose monitoring to predict hyperglycemic events by applying machine learning methods [39]. Using neural network models, it has been shown that hypertension, hyperglycemia, dyslipidemia, and a range of risk factors can be predicted from retinal fundus images obtained from a cross-sectional study of chronic diseases [176]. However, there is a pressing need to translate AI-based methods for supporting clinical decisions in different healthcare departments like emergency departments.

This section aimed to leverage data from the emergency department of the San Carlos Clinical Hospital in the Madrid region of Spain to predict hyperglycemic and hypoglycemic events that may occur within the hospital setting. Furthermore, a secondary objective of this research is to ascertain the predictive strength of each variable in terms of its significance in forecasting glycemic events.

4.3.2 Literature Review

Due to its rapidly increasing incidence supported by an unhealthy modern lifestyle, DM has been viewed by the United Nations and the World Health Organization (WHO) as a global epidemic, making it a major health concern [158]. Its early and accurate diagnosis can prevent its related complications. Nowadays, early diagnosis and prediction of diabetes and its related events are possible due to the development and application of AI and ML techniques, reducing the burden on medical professionals and patients. Over the years, numerous research studies focused on the implementation of AI and ML techniques to various aspects of the disease; from enhancing its diagnosis and prediction to helping in managing diabetes and glycaemic control, researchers have shown the ability of these techniques in enhancing and advancing the field of diabetes research. In this section, we provide a brief overview of the different applications of ML and AI in diabetes care.

Diagnosis and prediction of diabetes

Various algorithms and models are used for early detection, diagnosis, and prediction of diabetes. In one of the studies in this field [85], LR and DT were used to predict T2D and understand the risk factors related to T2D to help classify high-risk individuals and prevent, diagnose and manage diabetes. A dataset of 268 female diabetic patients with various medical attributes (nine attributes) was used, and their ML analysis found five main predictors of diabetes, including plasma glucose concentration, frequency of pregnancy, body mass index, age, and diabetes pedigree function, with a prediction accuracy of 78.26% and a cross-validation error rate of 21.74%. The same dataset with the same attributes was used in a similar study [139]. Their objective was to predict if a person has diabetes or not based on the diagnostic measurements of the patient, using six different ML algorithms: KNearest Neighbours (KNN), NB, SVM, DT, LR and RF. SVM and KNN provided an accuracy of 77% for predicting diabetes. The feature importance scores revealed plasma glucose concentration, body mass index, and age as the most important features in predicting diabetes. In another study [181], the same dataset of 268 female diabetic patients has been used to predict diabetes as well but contains only eight attributes compared to the previous studies, in addition to using a second dataset that includes both healthy people and diabetics that contains 14 features. Three algorithms were used: DT, RF, and neural network. According to their experimentation, using all features when predicting diabetes yielded better results, with the RF prediction reaching the highest accuracy among the three classifiers. However, using only the feature fasting glucose for the prediction had better performance, making it the most important attribute for prediction, but failed to achieve the best results, as there is a need for more attributes to make a better prediction.

Complications prediction

Predictive models can be used to predict the onset of diabetes-related complications and identify patients at higher risk of having complications. As an example, researchers [54] developed a robust and automated diagnostic technology using deep learning for diabetic retinopathy screening, a common complication of diabetes. The algorithm processed a total of 75137 publicly available fundus images from diabetic patients and classified them as healthy or abnormal (having diabetic retinopathy). The measurements reported an AUC of 0.97, with a sensitivity and specificity of 94% and 98%, respectively. The authors showed that their algorithm can accurately classify and identify healthy cases from the ones that need evaluation from an ophthalmologist, reducing the rate of vision loss.

In another study [84], several supervised ML algorithms, namely, LR, SVM, the CART DT, RF, AdaBoost, and XGBoost, were applied to predict and classify different complications related to diabetes, including metabolic syndrome, dyslipidemia, neuropathy, nephropathy, diabetic foot, hypertension, obesity, and retinopathy. Experimentation is done using a dataset of 884 records with 79 features. The outcomes revealed the RF, AdaBoost, and XGBoost models as best performers and total cholesterol, diabetes age, gender, BMI, and blood pressure as the most useful features for predicting complications.

Personalized treatment and management of diabetes

Given the ability of ML and AI techniques in processing and analyzing large amounts of patient data, providing tailored recommendations and developing personalized treatment plans can optimize diabetes management. AI technology can assist in self-management by helping diabetic patients in food intake tracking [105], physical activity tracking for disease monitoring [38], and even personalized food recommendations for people with a genetic predisposition to diabetes [167]. Furthermore, identifying whether an episode of hypo or hyperglycemia is likely to happen is of interest in diabetes therapy, as in the work led by Sudharsan [161] where models accurately predicted hypoglycemic events with a high degree of sensitivity and specificity. A recent review summarizes the ML approaches for diabetes care, with a focus on decisions on antihyperglycemic drug treatment prescriptions [51]. Glycaemic variability or fluctuations in blood glucose level is an indicator of diabetes management. In another study [133], ML models were developed using CGM data to predict short-term blood glucose levels in T1D. In addition, for controlling insulin therapy, authors of one study [180] suggest that ANN can potentially serve as an insulin dose controller in diabetics type 2.

4.3.3 Materials and Methods

Participants

We enrolled diabetic clinical samples from San Carlos Clinical Hospital in Madrid, Spain. We considered just patients over 18 years old with a diagnosis of type I or II diabetes who were admitted to the emergency department. The data collection for this study began in July 2018 and ended in July 2019. Table 4.8 shows the main characters of the participants in terms of age and gender. According to the table, most of the database information belonged to the older generation. However, the gender distribution was balanced between males and females.

The analyzed dataset contains 55 variables that provide various information in several domains. This retrospective multicenter analytical and observational study includes demographic, clinical, and analytical variables. In total, we had information related to 1415 case studies.

Clinical Measures and Data Collection

The variables were collected retrospectively. They consisted of qualitative and quantitative variables. Table 4.9 presents a comprehensive list of the variables employed in this study, their respective coverage areas, and a concise description of each. Additionally, the table includes the code utilized in the subsequent section to explicate the correlation between the variables.

All variables were collected by medical personnel belonging to the SEMES-Diabetes working group within the Spanish Society of Urgencies and Emergencies.

Table 4.8. Participants demographic characters

Age	Mean	70.35
	Median	70.35
	Variance	402.79
	SD	20.07
	Q1	59.00
	Q3	86.00
Sex	Female	697.00
	Male	718.00

4.3.4 Results

The subjects in this dataset are classified into two groups based on their glycemic event: hypoglycemia and hyperglycemia. Table 4.10 shows how each group is distributed in the dataset. In order to predict the glycemic event of the patients of the emergency section, the binary classification method was used, and ten machine learning algorithms were applied to the dataset. In the following sections, the importance of each variable and the performance of each ML algorithm will be evaluated.

Feature Importance

The purpose of this step was to identify the importance of predictors when it comes to predicting the target value. Information related to the ML models should be extracted and explained in a manner that is humanly understandable in order to explain how they work and make a decision. It is helpful to be aware of the critical predictors and the amount of contribution they provide. SHAP is one of the practical, instance-based explaining methods that can provide the mentioned explanations about the ML methods and the role of variables [109]. A Shapley value is an explanation method derived from coalitional game theory [147]. This is calculated by considering the average marginal contribution of one feature value across all possible coalitions [147]. SHAP provides a method to estimate Shapley values and more. In addition to explaining one instance, SHAP can also summarize model predictions by combining Shapley values of all

Table 4.9. Variables and their categories

MAIN CATEGORY	SUBCATEGORY	VARIABLES	Code	TRANSLATION
Demografic information	Visit date	f.at_urg	1D	Emergency care date
	Sex	Sex	2D	Male or female
	Age	Age	3D	Age of the patient
Clinical information	Assessment upon arrival to the emergency room	Recog_glu	1C	Blood glucose level upon arrival at Emergency Department
		TAS	2C	Systolic blood preasure
		TAD	3C	Dyastolic blood preasure
		Tª	4C	Temperature
		FC	5C	Heart rate
	Personal history	CI	6C	Ischemic heart disease
		IC	7C	Heart failure
		Enf.Vasc	8C	Peripheral arterial vascular disease
		ACVA	9C	Strokes
		HTA	10C	Arterial hypertension
		IREC	11C	Chronic renal failure
		DL	12C	Dyslipemia
		DM	13C	Diabetes Mellitus
		Tipo DM	14C	Type of DM
		Año DM	15C	Year of DM diagnosis
		Demencia	16C	Dementia
		EPOC	17C	Chronic Obstructive Pulmonary Diseas
	Degree of dependency and fragility	SB	18C	Basal situation according to activities of daily living
		Charlson	19C	Fragility according to score on Charlson scale
	Previous antidiabetic treatment	Tto_prev_DM	20C	antidiabetic treatment before to admission to the Emergency Department
		Metf_prev	21C	Previous Metformine treatment
		Sulfo_prev	22C	Previous Sulfonylureas Treatment
		Meglitini_prev	23C	Previous meglitinindas treatment
		TZD_prev	24C	Previous Tiazolinodiona Treatment
		aGLP1_prev	25C	Previous GLP1 analogues treatment
		SLGT2_prev	26C	Previous SGLT2 inhibitors treatment
		Comb_prev	27C	Previous Treatment in combination with various antidiabetics
		Basal_L_prev	28C	Previous treatment with long acting basal insulin
		Basal_UL_prev	29C	Previous treatment with ultra-long acting basal insulin
		Rápida_prev	30C	Previous treatment with short-acting insulin
		Ultrarrapida_prev	31C	Previous treatment with faster acting insulin
		NPH_prev	32C	Previous treatment with intermediate acting basal insuline
		Mixtas_prev	33C	Previous treatment with mixed insulin
		Terapia ins_prev	34C	Previous insulin therapy
	Antidiabetic treatment at hospital discharge	Tto_alta_DM	35C	Treatment of DM upon discharge from the Emergency department
		Metf_alta	36C	Treatment with metformine upon discharge
		Sulfo_alta	37C	Treatment with sulfonylureas upon discharge
		Meglitini_alta	38C	Treatment with meglitinidas upon discharge
		TZD_alta	39C	Treatment with tiazolinodione upon discharge
		iDPP4_alta	40C	Treatment with DPP4 inhibitor upon discharge
		iSGLT2_alta	41C	Treatment with SGLT2 inhibitors upon discharge
		Basal_L_alta	42C	Treatment with long acting basal insulin upon discharge
		Basal_UL_alta	43C	Treatment with ultra-long acting basal insulin upon discharge
		Rapida_alta	44C	Treatment with short acting insulin upon discharge
		Ultrarrapida_alta	45C	Treatment with faster-acting insulin upon discharge
		NPH_alta	46C	Treatment with intermediate acting basal insulin upon discharge
		Mixtas_alta	47C	Treatment with mixed insulin upon discharge
		Terapia ins_alta	48C	Insuline therapy upon discharge
Analytical information	Blood count	Leucos	1A	Leukocytes
		Plaquetas	2A	Platelers
	Biochemestry	Urea	3A	Urena
		FG	4A	Glomerular Filtration
		Cr	5A	Creatinine

Table 4.10. Distribution of the subjects.

Hyperglycemia	Hypoglycemia
430	985

instances [109]. Figure 4.7 is the SHAP beeswarm summary plot that shows the first 20 important variables and their effects on the predictions. Beeswarm plots provide an information-dense representation of how the top features in a dataset affect the outcomes of a model. A single dot on each feature row represents each instance. The dots "pile up" along each feature row, showing the density of a variable. Features are sorted along the y-axis by the sum of SHAP value magnitudes for all instances from top to bottom [109]. The distribution along the x-axis clearly displays the impact of each feature on the predictions of the model. Dots are colored according to the values of features: red for high values and blue for low values [109]. According to figure 4.7, the blood glucose level upon arrival is the most impactful feature. There is a threshold among variables. Moreover, the figure shows that eleven of the first 20 important features belong to the information related to the previous treatment of the patients.

The second method for understanding the role of each variable in predicting target value is plotting correlations. This method not only detects the association of the target value with all other predictors but also measures the correlation between all the variables. Figure 4.8 shows the described correlation between variables. According to the figure, there is a low correlation between variables in the dark blue part of the table, while moving toward the red part, there is an increase in the amount of correlation. The orange square sections show that these variables provide approximately the same information to the ML algorithms, and the algorithms can likely obtain the same result even if some of these variables were deleted. In our case, these variables are related to the kind of previous treatment and the kind of treatment upon discharge. As a result, the correlation between the mentioned two groups of variables and the target value "Glycemia" showed that they were able to guide the ML methods to reach better results. Results from current data prove that the effects of antidiabetic treatments on glycemia events in diabetics are an important factor in preventing or reducing such events.

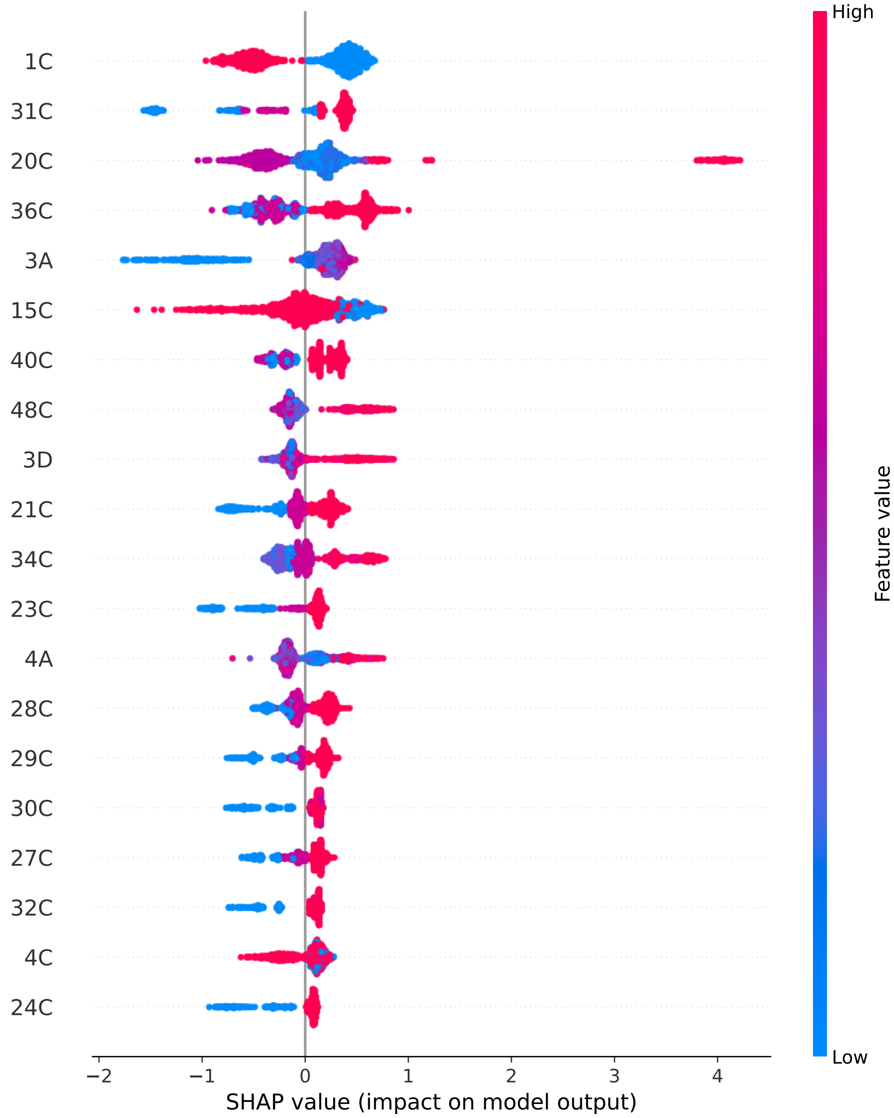

Figure 4.7. Importance of the predictors

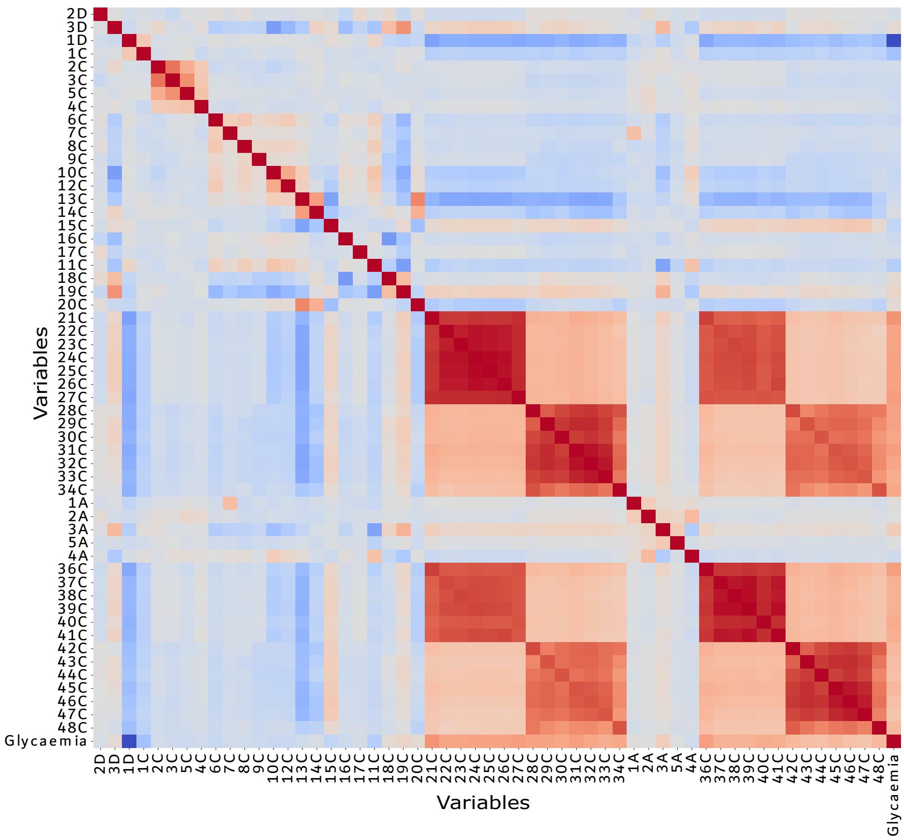

Figure 4.8. Correlation between the variables.

Predictions

In the following section, we present the results based on the dataset previously described. This paper shows how ML methods can provide valuable insights and predictions about the glycemic events of emergency section patients due to the fact that our database comprises two groups of participants: those with hypoglycemia and those with hyperglycemia. A total of ten ML algorithms were applied to the dataset in order to evaluate and compare their performance. Table 4.11 shows the result of each algorithm. Based on the results, we show that ML methods can accurately predict the glycemia event of the patients in the emergency section with up to 95 percent. The best result was obtained by gradient boosting, while the worst performance was achieved by stochastic gradient descent with 90% accuracy. Regarding accuracy, ada boosting and MLP achieved one percentage less than the best ML method with 94%. When it comes to comparing the whole performance of all ten ML algorithms, it can be seen that all algorithms in terms of all four performance metrics can provide results among 90% to 95%.

Table 4.11. The result of all used ML methods.

	PRECISION	RECALL	F1_SCORE	ACCURACY
RANDOM FOREST	0.95	0.94	0.94	0.95
SVM	0.93	0.92	0.91	0.92
K Nearest Neighbors	0.92	0.92	0.91	0.92
Decision Tree	0.92	0.92	0.92	0.92
Logistic Regression	0.91	0.91	0.91	0.91
MLP	0.94	0.93	0.93	0.94
Stochastic Gradient Descent	0.92	0.90	0.90	0.90
Ada Boost	0.94	0.94	0.94	0.94
Gradient Boosting	0.95	0.95	0.95	0.95
XGradient Boosting	0.95	0.95	0.95	0.95

Figure 4.9 provides a visualization of the Receiver Operating Characteristic (ROC) curves obtained based on the results of ten different ML classifiers. As explained in chapter two of this thesis, a curve displaying the true positive and negative rates for the classifier under various operating conditions. Depending on the specific purpose of the statistical analysis in question, an analysis of this curve can be used to determine the balance between true positives and false negatives. This analysis uses 80 percent of the dataset for training, while 20 percent of the dataset is used to evaluate the performance of each classifier. In order to

obtain a more comprehensive evaluation of the performance of the classifiers, we evaluated their accuracy using the AUC. Based on the figure, it can be seen that the results are consistent with those shown in Table 4.11. According to the figure, the best performance belong to the gradient boosting classifier, xgradient boosting classifier, RF, and SVM.

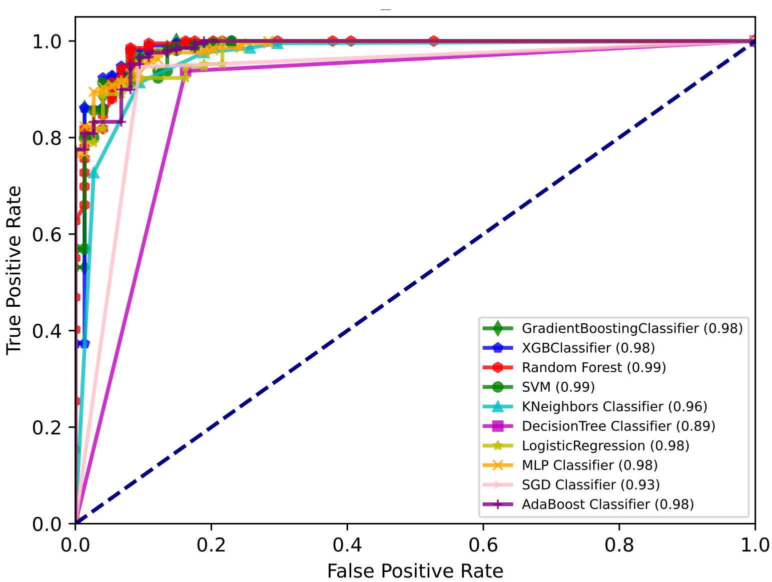

Figure 4.9. ROC curves of different ML models (AUC values).

The classification threshold is also known as the threshold of the decision in ML. It is a method for clearly differentiating between the two classes. The threshold must be reliable to be used in the classifier. To achieve the best results, it is essential to adjust the classification thresholds for each ML application individually. Although no perfect machine learning model exists, we can make smart decisions to optimize performance. One option that can be considered in the context of optimization is to adjust the classification thresholds. The threshold plays an important role in limiting misclassifications, so setting the right classification threshold is extremely important. Figures 4.10, 4.12, and 4.11 show the precision, recall, and f1 score of the three best ML methods in the case of using different thresholds. According to the figure, gradient boosting and xgradient boosting perform better than random forest in contexts where the recall is considered. Analyzing the graphs at the peak of the F1-score curve shows the recall belonging to the gradient boosting (98%) achieves higher values than the value of

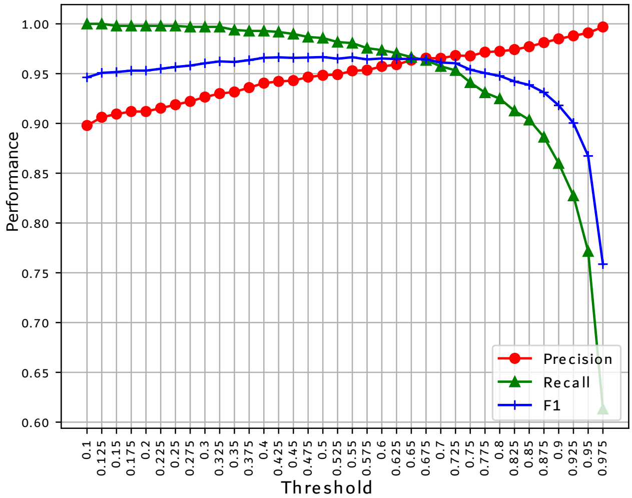

Figure 4.10. Threshold of the Gradient Boosting Classifier

recall obtained by random forest (96%) and xgradient boosting (97%). It is worth noting that at the highest point of the F1-score curve, all three ML methods had a Precision performance of over 90%. Based on the information provided by the figure, gradient boosting and xgradient boosting can be considered as the best supervised ML models for our purposes. Moreover, regarding the classification threshold, the best values are 0.575 for gradient boosting, 0.475 for xgradient boosting, and 0.55 for RF.

Another method for improving the performance of ML methods and understanding their behavior in different conditions is setting different weights for each class. Based on the weights assigned to the two classes, figure 4.13 represents the precision, recall, and F1 score for the RF algorithm as one of the best ML methods. According to the figure, the highest performance (94%) of the algorithm in terms of precision is when the weights of the classes are one for the negative class and four for the positive class. Additionally, using the same weights (Negative: one and Positive: four), the f1 score is reached at its best. On the other hand, the best results regarding recall occurred when class weight was set at one and three. It should be mentioned that the best performance of all three performance metrics is considered in the selected class weight. Therefore, the chosen class weights were positive: six and negative: one.

Following the application of ML classifiers and knowing their performance, a

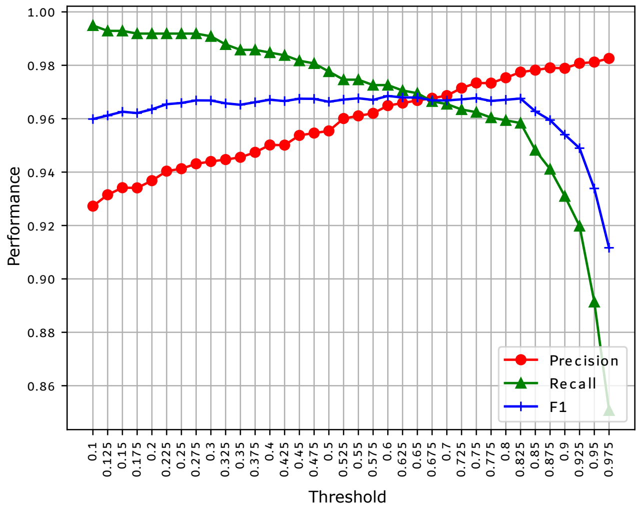

Figure 4.11. Threshold of the XGradient Boosting Classifier

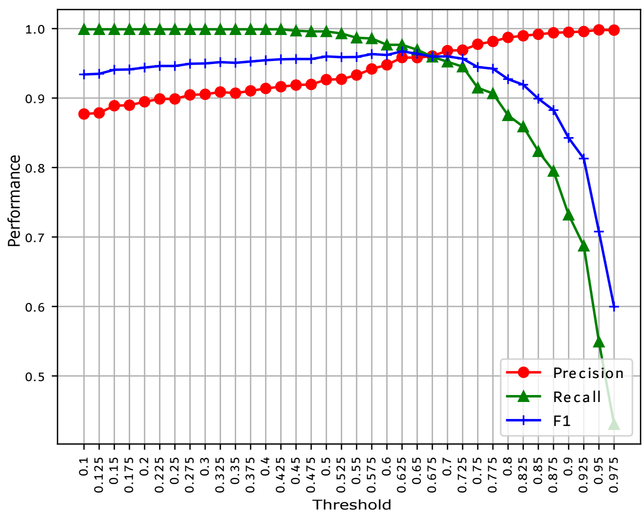

Figure 4.12. Threshold of the Random Forest Classifier

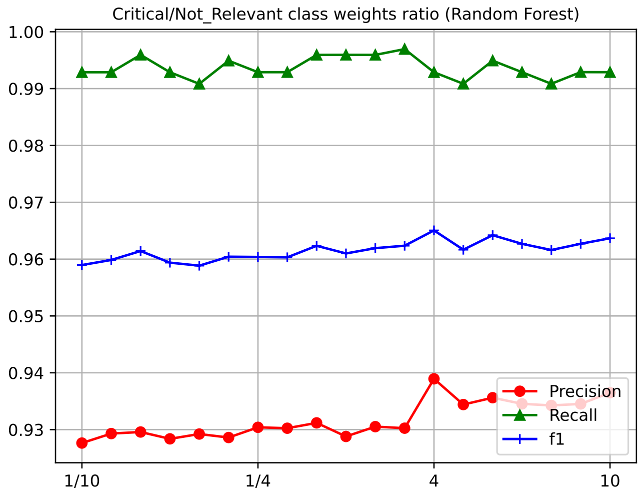

Figure 4.13. Performance of random forest with different class weights

further analysis has been performed to identify which type of cases will improve performance. In order to investigate the causes of performance, we only considered the gradient boosting model. In the following, there are two kinds of plots. One group is reporting Cumulative Distribution Function (CDF) of the five most significant features, and the second group is Partial Dependence Plots (PDP) of the same group of features. PDP provides a visual representation of the effect of one feature of interest on the prediction results while marginalizing the implications of the other features. Partial dependence is the expected target response as a function of a feature of interest. The grey bars on the plot indicate the data distribution, and by using the blue line, the expected performance has been shown.
Regarding the CDF plots, orange and red colors are used to indicate the cases predicted correctly, while blue and green show the instances incorrectly classified. Figure 4.14 shows the CDF and PDP of the five most critical features. According to the information provided by all plots, including CDFs and PDPs, the features of interest are not independent of the rest of the features. Therefore, based on the CDF plots, the incorrect and correct predictions happened almost in all values of each feature, as the changes in the value of each feature of interest do not affect the prediction result significantly in the PDP plots. In summary, the obtained results are based on the contribution of all variables.

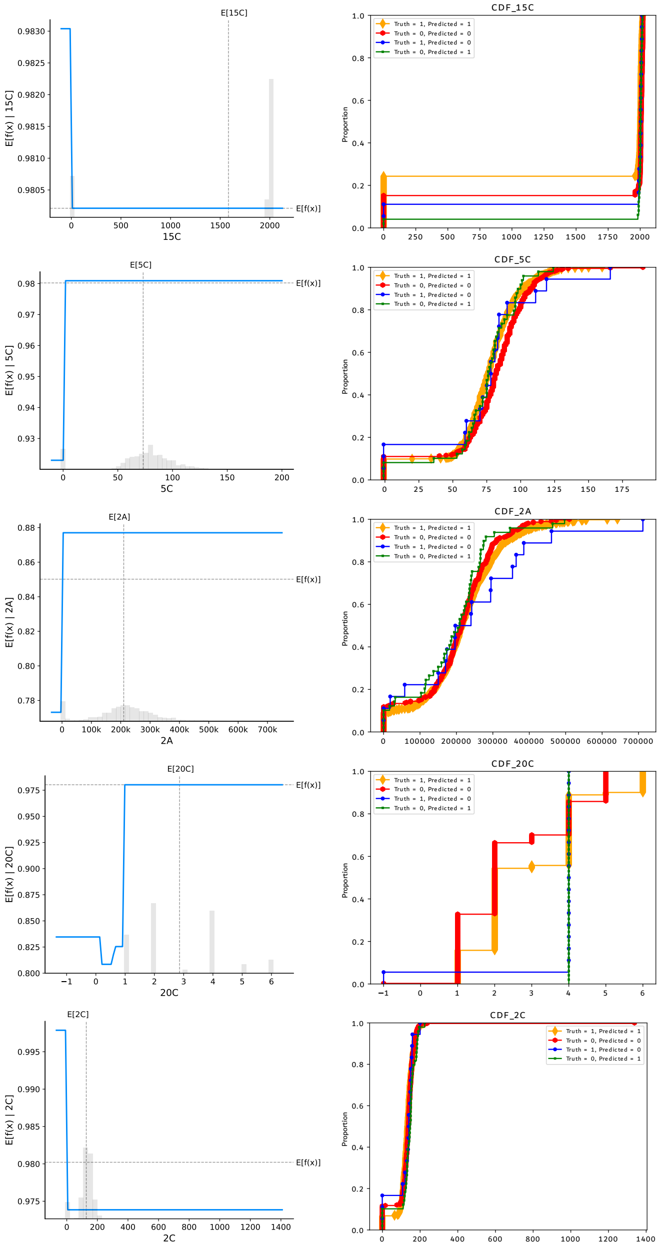

Figure 4.14. Role of the first five important features

4.3.5 Discussion

In the present study, we collected three different groups of information from two groups of emergency section patients of San Carlos Clinical Hospital in Madrid, Spain, who experienced hypoglycemia or hyperglycemia, intending to detect the role of mentioned variables in the glycemic event of the diabetic patients and predict hypo/hyperglycemia based on those predictors and by using different ML methods. we used ten different ML algorithms to reach this goal. Moreover, we work on understanding the role of different variables and ranking them based on their importance by applying two different methods. we set up the methodology based on the assumed targets. In the first step, the input data was processed carefully by cleaning and normalizing them, then we started to analyze the dataset.

The current model allowed us to determine the role of each predictor. Therefore, it can illustrate the interplay between different variables, including clinical, demographic, and analytical variables. Moreover, the result proves that glycemic events like hypo or hyper-glycemia can be predicted by taking into account the mentioned data related to the patients, and no single predictor can be used for this task. As a result of analyzing the role of each predictor, the data related to the previous treatments were among the best predictors for predicting the glycemic event and had a high correlation with it.

Despite the prognostic potential of the described methodology and the opportunity to uncover unknown relations between different kinds of meaningful variables. Therefore, the information obtained from the present study can be used to set up guidelines and policies for treating hospitalized patients in the emergency section. Moreover, the findings prove the potential of AI methods to be employed in healthcare 4.0. From this perspective, this study might pave the way for future investigation aiming to practically apply the described method to reduce the risk of glycemic events, manage diabetes in a more feasible way, and, subsequently, diminish the rate of mortality for this reason. Such a system not only allows physicians and clinicians to select medicines and treatment methods wisely but also prevents glycemic events in advance before they happen. Moreover, the result can also help to use the most appropriate indicator for the estimation of the risk profile of each individual and accordingly provide a personalized and well-timed treatment.

4.3.6 Conclusion

In conclusion, here we present an ML approach to identify and understand variables that are suitable to predict glycemic events and support medical doctors in their decisions. In the current section, we used three groups of information, including demographic, clinical, and analytical information on two diabetes groups.

The majority of the data that was used belonged to the older generation. We showed by using ML methods, we cannot only identify the role of predictors and understand them but also predict the glycemic event of the patients in advance with an accuracy of up to 95%. We showed that clinical information, in particular data related to the antidiabetic treatment method, is critical for predicting hypo/hyperglycemia events. Two methods, namely SHAP and a correlation plot, were employed to obtain the mentioned results. Future studies should confirm our preliminary findings, and this model can be applied to different datasets, including healthy and diabetic subjects.Additionally, this model can be utilized for the design and development of healthcare systems such as comprehensive clinical AI-DSS by health authorities to implement public health programs and provide better smart healthcare services.

Chapter 5

Cardiology

5.1 Introduction

Cardiovascular diseases are responsible for a significant number of deaths globally [21]. Cardiac arrhythmias are one of the most common causes of cardiovascular disease, in which the heartbeats differ from standard patterns [153]. The most common method of diagnosing and following up on the situation of patients is the analysis of critical portions of their ECGs due to its non-invasive, inexpensive, and safe nature [23]. An ECG signal records the electrical activity of the heart with the help of electrodes placed on the body [114]. It is a non-invasive device used to record the direction and magnitude of the electrical activity of the heart that occurs as a result of the depolarization and repolarization of the atria and ventricles [87]. It is important to note that ECG readings can vary significantly among individuals due to differences in heart mass orientation, position, size, anatomy, the conductivity of cardiac muscles, activation order, chest configuration, age, sex, relative body weight, and other electrophysiological factors. These variations directly impact the morphology, amplitude, and time intervals of the fiducial points of the heartbeat, and as a result, ECG is a highly unique biometric system. Therefore, it is crucial to consider these differences when analyzing ECG readings for diagnostic purposes or biometric identification [87]. The accuracy of segmenting the ECG beats and detecting fiducial points are pivotal for analyzing ECG signals and knowing certain cardiac conditions [24]. Clinical practitioners currently analyze the ECG signals manually to determine these cardiac conditions [71]. It is, however, a time-consuming and skill-intensive process to interpret data manually [153]. Incorrect interpretation of ECG signals may result in incorrect clinical decisions and may endanger the life and health of patients. As ECG technology has developed rapidly and as the number of cardiologists is insufficient, accurate and automatic diagnosis of ECG signals has become an interesting research topic for many scientists [153].

The identification of the onsets and offsets of QRS complexes, as well as P and T waves in an ECG signal, is known as segmentation or delineation [114]. As part of the ECG analysis, QRS complexes P and T waves are detected, and then their shapes, amplitudes, relative positions, and other characteristics are analyzed. Due to the following reasons, automated ECG segmentation is challenging. The high variability in the shape of the QRS complex, the reduced amplitude of the P-wave, and smooth transitions at the beginning and end of the T-wave make it difficult to determine the exact position of the fiducial points. Furthermore, the lack of a universal consensus on the definition of these points further complicates the task [23]. Also, some cardiac cycles may not contain all the standard segments, such as the P wave [114].

Various techniques are available in literature for automatically segmenting ECG signals. In the work led by Aspuru, the authors used linear regression to segment ECG signal to detect R points of it [13]. They used 260 ECG signal to evaluate the detection approach performance with a sensitivity of 97.5% for Q-point and 100% for the rest of ECG peaks. In another work, the authors used machine learning methods for the automated detection of arrhythmias in ECG segments. They mentioned in their paper that random forest was the best ML method [127]. In the study conducted by Jikuo Wang et al., [169] a new Convolutional Neural Network (CNN) with a non-local convolutional block attention module has been proposed to classify ECG heartbeats automatically. Using the MIT-BIH arrhythmia database, the proposed method achieves an average F1 score of 0.9664 and an AUC of 0.9314 using the PTB-XL ECG database.

Detecting heart disease by analyzing and processing recorded signals is a continuous process that is prone to human errors. To address this issue, system classification is crucial for automatic identification, which has garnered significant attention from researchers [112]. Following the current literature in this field regarding the classification of ECG signals. In this study, we evaluated several AI methods on the PhysioNet's QT public dataset in order to find the most effective method for automating the detection ofECG signals parameters. We tested six different methods of classifying ECG signals provided in the PhysioNet's QT dataset.

5.2 Literature Review

The ECG signal segmentation has witnessed significant attention and advancements in recent years, fueled by the imperative need for accurate cardiac health assessment. As a fundamental diagnostic tool in clinical practice, the

ECG provides a wealth of information embedded in its temporal dynamics. Efficiently extracting and interpreting this information hinges on the precision of signal segmentation, a process crucial for delineating distinct cardiac events such as the P, QRS, and T waves. Shedding light on the diverse approaches and innovations that have shaped the current state of ECG signal segmentation. The work led by Maglaveras [104] conducted a comprehensive review of ECG pattern recognition and classification using non-linear transformations and Neural Network (NN). Their review encompassed non-linear transformations of the ECG, the application of principal component analysis (both linear and non-linear), methodologies for mapping transformed data into n-dimensional spaces, and the utilization of NN-based techniques for ECG pattern recognition and classification. The specific challenges they addressed included the recognition and classification of QRS/PVC (Premature Ventricular Contraction), identifying ischemic beats and episodes, and detecting atrial fibrillation. Furthermore, they presented a generalized approach to classification problems in n-dimensional spaces, incorporating techniques such as NNs, Radial Basis Function Networks (RBFN), and Non-Linear Principal Component Analysis (NLPCA). The study also evaluated the performance measures, including sensitivity and specificity, of these algorithms. In another work, the researchers [11] have presented an innovative Hidden Markov Model (HMM) approach for online beat segmentation and classification of ECG. The HMM framework was chosen for its adeptness in beat detection, segmentation, and classification, making it well-suited for ECG analysis. The study explored novel aspects in HMM-related research, including waveform modeling, multichannel beat segmentation, classification, and unsupervised adaptation to the ECG of patients. The evaluation was conducted on the two-channel QT database. The approach exhibits beat detection performance with a sensitivity of 99.79% and a positive predictivity of 99.96%, using a test set of 59 recordings. Additionally, an original classification strategy enables the detection of premature ventricular contraction beats. Beraza and her colleague [24] have focused on evaluating nine ECG segmentation algorithms selected from the literature using a standardized protocol and a dataset of 100 signals from the PhysioNet's QT database. The aim was to compare their performance for accurately identifying fiducial points within an ECG. Notably, algorithms based on discrete wavelet transform achieved 100% sensitivity for QRS complex onset and offset. Another algorithm utilizing Multi-scale Morphological Derivate demonstrated high sensitivities (99.81% , 98.17% , and 99.56%) for P-wave peak, onset, and offset detection, respectively. For T-wave segmentation, an algorithm based on the Phasor transform performed well with sensitivities of 97.78%, 97.81%, and 95.43% for peak, onset, and offset detection. Hedayat [5] introduced a method for segmenting ECG signals for predicting heart symptoms and medication effects. The proposed method employs a Recurrent Neural Network (RNN) with LSTM layers, classifying each ECG sample into P-wave, QRS-wave, T-wave, or neutral (others). The study showed that DL sequence learning methods, particularly using simple local features, outperformed traditional Markov models in terms of accuracy. Notably, on T-wave segmentation, the approach achieves 90% accuracy compared to the 74.2% accuracy achieved by Markov models. In another work, the researchers [140] have presented and validated a model-based Bayesian framework for ECG beat segmentation. The method employed a modified Extended Kalman Filter (EKF) algorithm, integrating parameters of the ECG dynamical model. AutoRegressive (AR) model was introduced for each dynamic parameter, enabling the estimation of hidden-state variables for analytic Fiducial Point (FP) extraction. The filter was applied to various ECG signals, showing its effectiveness in filtering and tracking parameters of a Gaussian Mixture Model (GMM) for the ECG signal. The EKF structure not only estimates the clean ECG but also the parameters of the model, allowing the matching of Gaussian kernel parameters to the locations of fiducial points (FPs). The mathematical model efficiently tracks morphology and derives rules for locating characteristic wave peaks based on estimated state variables. Adaptive Kalman Filters (KFs) are viewed as adaptive filters, ensuring reliable estimations and precise FP extraction. The method was validated using various ECG recordings, showing reliable FP detection with high specificity (99.83%) and positive predictivity (99.98%). Another work [14] addresses the need for improved processing algorithms for ECG signals, emphasizing aspects like processing speed and efficient detection of all ECG peaks. The proposed method utilized parabolic wave functions to segment ECG signals and extracted valuable information for discriminating between normal and abnormal heartbeats. The approach implements a simple LR process and analyzes 260 ECG signals, achieving an average sensitivity of 99.5% for identifying all peaks. To validate the approach, an ECG sensor was developed to acquire real-time signals, visually represented based on the implementation of the detection algorithm. Despite a high average sensitivity, there was a 3% error rate in correctly detecting the Q point. It should be mentioned that the dataset used for training and testing had limited examples of non-sinusal or abnormal beats. in the paper led by Aman Malali [106], the authors introduced a Convolutional Long Short-Term Memory (ConvLSTM) neural network for the segmentation of ECG waveforms. The proposed model, comprising a convolutional layer and a Bidirectional LSTM architecture, utilizes additional features like the derivative and smoothened ECG wave to achieve segmentation. The ConvLSTM model showed superior performance compared to traditional Markov models in segmenting ECG waves, providing a valuable tool for ECG analysis and diagnosis. Aboli N., et al [102] introduced a new approach for segmenting ECG waves, termed semantic segmentation, commonly employed in image segmentation. The proposed method includes a hybrid channel-mix convolutional and bidirectional LSTM architecture, addressing temporal dependencies and lead variability. The model, implemented on the QT database, had an accuracy ranging from 96% to 98.56% . The average and weighted accuracies were 96.72% and 95.54%, respectively. Another interesting work [7] suggested using a Multi Hidden Markov Model (MultiHMM) for ECG beat segmentation. The method was evaluated using the Physionet QT database and a Swine ECG database, comparing it with Classic HMM and a method based on Partially Collapsed Gibbs Sampler (PCGS). Results indicate that MultiHMM outperforms other benchmark methods in terms of Root Mean Square Error (RMSE) values and demonstrates smaller error variability. The researchers in another work proposed an attention-based Convolutional Bidirectional Long Short Term Memory (Conv-BiLSTM) architecture network, leveraging local beat features and temporal sequencing to correlate ECG beats across different positions. The segmentation tool comprises three main blocks: Bidirectional Long Short-Term Memory (BiLSTM) for temporal context, CNN for spatial features, and an attention-based mechanism for abstractive representation. The performance of the ECG segmentation tool was evaluated in terms of accuracy, achieving 95% for ECG segmentation and 99.41% for FP detection [70].

5.3 Materials and Methodology

5.3.1 Database

The two-channel PhysioNet's QT database [93] was utilized to evaluate proposed DL models. In order to explain the dataset in more detail, it should be highlighted that the dataset includes 105 ECG recordings selected from various databases, showcasing a diverse range of QRS, T-wave, and P-wave morphologies. In each record, at least 30 beats were manually annotated by an expert. Furthermore, 11 of the recordings were annotated by a second expert. This database contains annotations on the peak, onset, and offset of the QRS complex (QRS-peak, QRSon, QRSoff) and the P-wave (Ppeak, Pon, Poff), as well as the peak and offset of the T-wave (Tpeak, Toff). Only some records (totaling 1345 beats) contain the onset of the T-wave (Ton).

Out of 105 records, 94 were used, and 11 were excluded during the prepossessing process due to annotation issues; some annotations were unavailable for these records.

5.3.2 Pre-processing

As an initial step in this research, the recorded files and their corresponding annotations were examined by visualizing the signal plots to facilitate a deeper understanding of the dataset. This allowed for a more comprehensive evaluation of the information at hand and paved the way for further analysis. In the subsequent phase of the experiment, the signals were subjected to a filtering process utilizing two techniques, namely wavelet decomposition and Waverec [69]. This approach was implemented to enhance the quality of the signals and to minimize the presence of noise. The wavelet decomposition process involves breaking down the signals into various sub-bands, each representing different frequency ranges, and filtering out unwanted frequencies [69]. The Waverec method, on the other hand, utilizes a mathematical algorithm to reconstruct the filtered signals from the sub-bands [69]. Overall, this approach was adopted to improve the accuracy and reliability of the results obtained from the analysis. Once the extraneous noise had been removed from the signals, the segments were divided into portions of equal size. The objective was to ensure that each segment shared the same amount of data, thereby making it easier to compare them and ensuring uniformity. Through such standardization of the signal data, any subsequent analysis or interpretation of the signals can be enhanced for improved reliability and accuracy. During preprocessing, the subsequent step entailed normalizing the complete dataset using the z-score normalization method [69]. This involved applying a statistical technique that adjusted the values of the variables to a common scale, bringing them to a more comparable range. The normalization of the dataset is a critical step in preparing the data for further analysis, as it ensures that the variables are not biased by their original scale or measurement unit. To implement the DL models, our approach involved the incorporation of three-dimensional NumPy arrays. These arrays were intended to serve as input data for the system. Then, we performed a series of merge operations to achieve a more balanced dataset and furnish the model with a well-balanced input. Specifically, we merged the "P-wave" and "PQ-wave" to form a single entity. Furthermore, we combined the "QR-wave", "RS-wave", and "ST-wave" into a unified entity. Finally, we merged the "T-wave" and "TP-wave" to create a cohesive entity, which resulted in a more balanced dataset with significantly improved accuracy. Table 5.1 includes a summary of the methods used to create a balanced dataset,

Table 5.1. Distribution of input

P, PQ	QR, RS, ST	T, TP
PQ	QRST	TP
89986	80982	148832

along with the distribution of each class. This information is vital in gaining a deeper understanding of the composition of the dataset and knowing whether it is a balanced database, as it can significantly impact the accuracy of predictive models.

5.3.3 Method of Obtained Result

Our approach entailed the utilization of Recurrent Neural Networks (RNNs) to segment the entire ECG beat to facilitate the modeling of the event sequence. This approach was chosen due to the inherent ability of RNNs to develop a deep understanding of sequential data by processing the inputs [69]. By adopting this approach, we were able to achieve a high degree of accuracy in the segmentation of ECG beats, enabling more precise analysis and interpretation of ECG data.

In response to the long-term dependency issue in traditional RNNs, among these architectures, LSTM was developed to overcome the problem [79]. The LSTM architecture allows information to flow through the network over extended periods, thereby avoiding the vanishing gradient problem observed in conventional RNNs [150]. We have developed a Conv-BiLSTM architecture to extract local and time-variant features [69]. This architecture is trained on data segments that capture information from past and future contexts, enabling the model to learn complex temporal patterns and improve its performance in capturing the data dynamics. Our approach involves applying a one-dimensional convolutional layer that convolves the signal x_{1d} with the kernel initialized by the 'glorot-uniform' method in the temporal dimension. The kernel weights W_{1d} are updated after each iteration to ensure optimal performance. These can be represented as:

$$y_{1d} = [W_{1d} \otimes x_{1d} + b]$$

Moreover, implementing a Conv-BiLSTM network has been undertaken to facilitate the training of temporal context in both directions of time. This approach enables the signal to be trained based on the temporal context, thereby enhancing the accuracy and effectiveness of the training process. Moreover, it enables the network to better capture the long-term dependencies and temporal correlations within the data. This can be expressed as follows (It should be mentioned that $\vec{h}_t$ and $\overleftarrow{h}_t$ are forward and backward LSTM outputs at time t):

$$h_t = Dropout\left(\left[LSTM\left(y_{1d_t}, \vec{h}_{t-1}\right), LSTM\left(y_{1d_t}, \overleftarrow{h}_{t+1}\right)\right]\right)$$

Additionally, we used the self-attention block. This study describes a block aimed at temporally correlating ECG sequences to compute abstract representations as a function of bidirectional temporal context. Correlations between the

ECG sequences can be used to represent attention weights. A bidirectional temporal context of ECG sequences can be leveraged in this method for extracting meaningful information and enhancing the accuracy of ECG-based applications. It should be clear that dot (h_{t_i}, h_{t_j}) is the attention score, which is the dot product of LSTM output by itself; J is the total number of LSTM outputs, σ_{max} the softmax function representing the attention weights. Weights for attention can be calculated as follows:

$$\sigma_{max}(z)_i = \frac{e^{dot\left(h_{t_i}, h_{t_i}\right)}}{\sum_{j=1}^{J} e^{dot\left(h_{t_j}, h_{t_i}\right)}}$$

Also, the dense layer with the ReLU activation method was used to introduce the feature of nonlinearity to a DL model and resolve the issue of vanishing gradients.

In order to evaluate the performance of the employed methods, a five-fold cross-validation procedure was implemented. A confusion matrix was utilized to obtain the resulting evaluation metrics. Figure 5.1 illustrates the methods used to achieve the results and what was utilized. The present study aimed to enhance

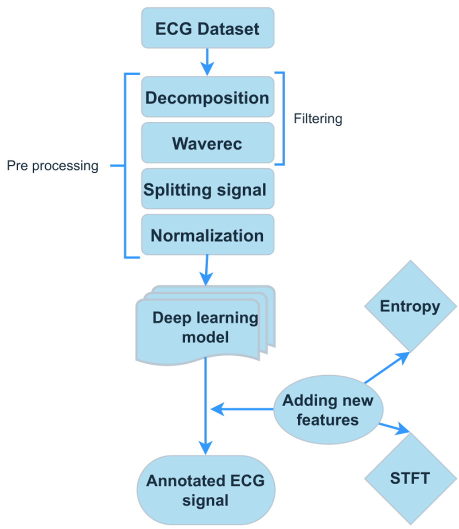

Figure 5.1. Utilized items in the study

the performance of DL method in the classification of ECG signals. Six different experiments were conducted to achieve this. The ensuing section presents the obtained results. The results of the experiments provide valuable insights into the efficacy of various approaches in improving the accuracy of the DL methods for ECG signal classification. These findings could have significant implications for the design and development of more precise and reliable ECG-based diagnosis systems. A technique utilizing a random seed was employed to ensure consistency between the training and test sets. The purpose of this technique was to maintain the integrity of the data and reduce the probability of errors or bias in the results. By implementing this approach, we were able to establish a standardized process for selecting and dividing the data into respective sets.

5.4 Result

As explained in the previous section, several methods have been developed in order to achieve better results for classifying ECG signals in three different classes mentioned in Table 5.1.

As an initial model, we defined a dense neural network layer that had 32 neurons, and we applied L2 regularization to the weights of the first layer. Then, we decided to consider an BiLSTM [142] with 32 units that are capable of generating output sequences based on the input sequences provided. In order to prevent overfitting 20% of the outputs of neurons randomly was set to zero. Then, we considered improving training speed, stability, and generalization performance by adding a "BatchNormalization" layer. To improve the generalization ability, Non-linearity, and control overfitting of the model, we defined a dense layer with 64 neurons, used ReLU activation, and applied L2 regularization to the weights with a regularization strength of 0.001. The forthcoming three layers were constructed by employing the last three layers. Then, in order to classify output based on probabilities, we used a dense output layer with dimout neurons and a softmax activation function. Finally, the model complied with the Adam optimizer by using categorical cross-entropy as the loss function and accuracy as the evaluation metric during training. By using the described model, we obtained an accuracy of up to 88%. Figure 5.2 displays the results by showing the accuracy obtained for each class using five-fold cross-validation. Based on the figure, the best performance belongs to the class called 'TP' with up to 91%, when the performance related to the 'QRST', was between 75% and 86%.

We built the next model by increasing the number of layers, duplicating the same layers, and changing some parameters. Therefore, we had three dense neural network layers, one with 32 neurons and the others with 64 neurons, and 'ReLU' as the activation method. Also, we had three BiLSTM layers with the same configuration as the first model. Further, two layers were added to protect against overfitting by setting 20% of outputs to zero. After applying all the

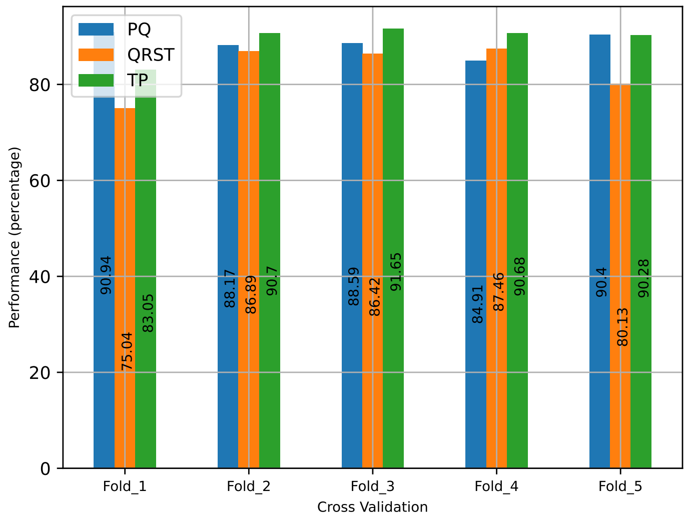

Figure 5.2. First model performance

mentioned changes, we used the model to classify the ECG signal. Figure 5.3 shows the obtained results for all five-fold cross-validation. According to the figure, the performance related to all three classes improved slightly. The best performance for the 'QRST' class reached 91%, and for two other labels, we obtained 93% of accuracy. Upon duplication and increasing the size of the model, the performance of the developed method exhibited a slight improvement. This result suggests that duplicating the layers may be a viable technique for enhancing the performance of the model.

In order to assess the efficacy of the adopted technique in optimizing the performance of the model, the size of the model was augmented by duplicating the existing layers. This was done to determine the extent to which the method can be employed to enhance the overall performance of the model. Figure 5.4 showcases the performance of the utilized model for each class when employing the five-fold cross-validation technique. The results indicate that increasing the model size, despite the additional computation requirements, leads to a higher degree of performance. Notably, the analysis reveals that the best performance levels achieved for the 'PQ', 'QRST', and 'TP' classes were approximately 96%, 90%, and 96%, respectively. Moreover, the model was able to attain an accuracy level of up to 90%. The outcomes of this part provide insights about the impact of the model size on its overall performance.

In the subsequent step, we evaluated the effectiveness of the model when

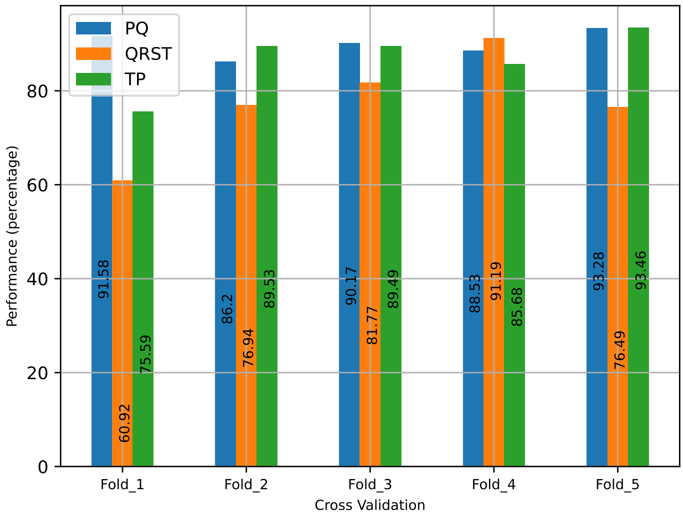

Figure 5.3. Performance of the second used model

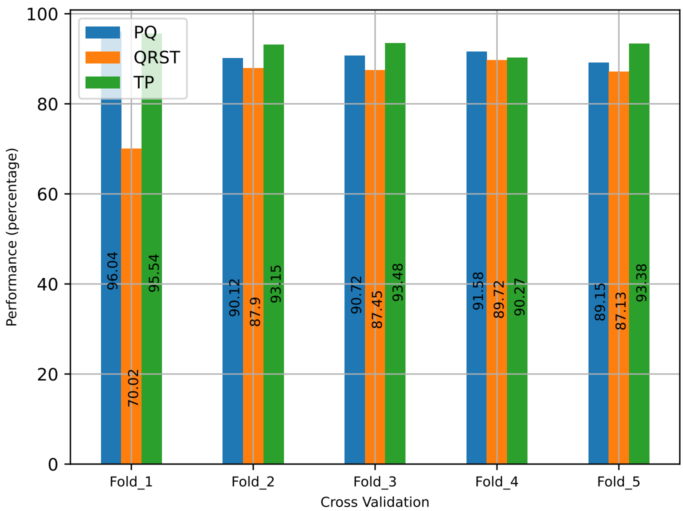

Figure 5.4. Performance of the third used model

the approximate entropy was computed. The resultant entropy values were then incorporated as a new feature into the dataset. Entropy measures serve as a valuable tool to gain insight into the intricacy or regularity of ECG signals. These measures can be used to quantify the degree of disorder or complexity of ECG signals, providing a meaningful metric for evaluating their quality by providing a more comprehensive understanding of ECG signals. In particular, approximate entropy is a statistical measure employed to quantify the degree of regularity or predictability of a time-series dataset. The development of this measure was aimed at assessing the complexity or randomness inherent in datasets, especially in the context of physiological signals such as ECG. Figure 5.5 depicts the performance of the same model when approximate entropy was added as the one extra feature. Based on the result provided in the figure, the methodology used did not improve the performance of the model in classifying ECG signal into described three categories. Overall, the results were the same as those of the previous method. For the first category ('PQ'), it reached up to 97%, and for 'QRST' and 'TP', the best performance was up to 93% and 96%, respectively.

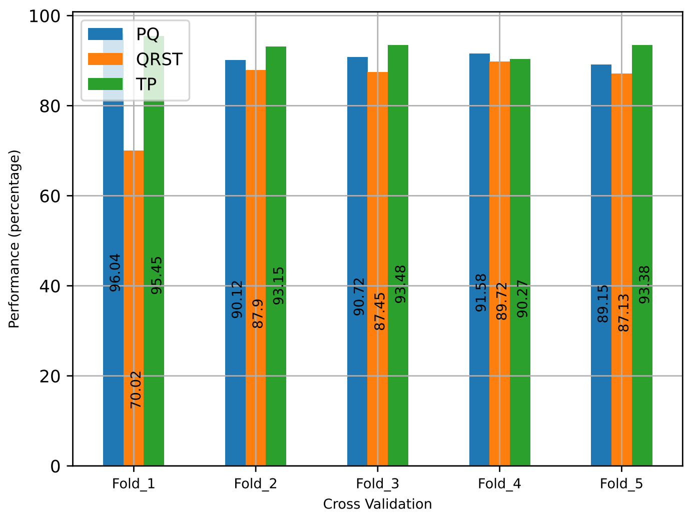

Figure 5.5. Performance of the model when entropy was added

In order to enhance the performance of classifying ECG signals, a specific step was taken. This step involved removing the values associated with approximate entropy from the dataset and adding the values related to the STFT as a new feature into the dataset. STFT is the technique employed to analyze the frequency content of a signal over time. The STFT is widely used in various fields, including audio and speech processing, image processing, and biomedical signal analysis. In analyzing the ECG signal, the STFT can be used to examine the frequency components present in an ECG signal at different time intervals. The same model, which was described previously, was used to classify the signal. Figure 5.6 shows the results when the described method was employed for our classification target. The presented diagram shows that the incorporation of STFT yields a slight enhancement in the performance of the method. This finding is noteworthy in the context of the method's overall capabilities and its potential applications in diverse settings.

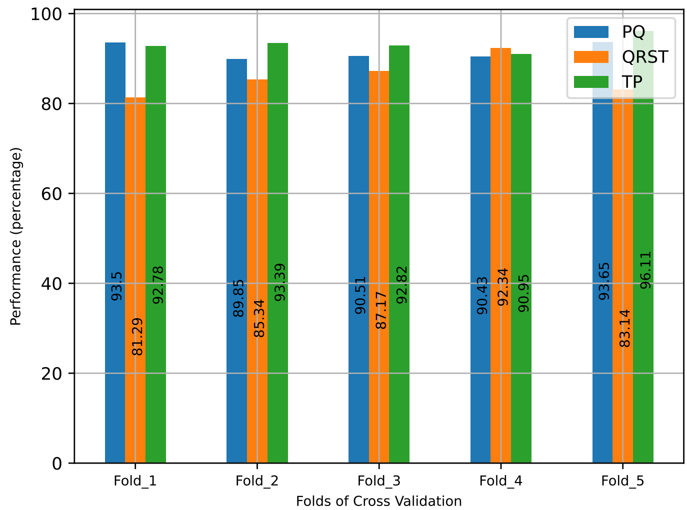

Figure 5.6. Performance of the model when STFT was added

In the last step, we added both STFT and approximate entropy to the data set, in order to check the power of collaboration of these two features of signal to improve the performance of the classification task. Using the same model, we evaluated how well the method classified ECG signals into the three categories mentioned. Figure 5.7 obtained results for each class in the case of five-fold cross-validation when the described methodology was employed. The obtained results indicate that although the average accuracy has increased to approximately 92%, it seems that the progress of each class has improved, albeit at a sluggish rate. Figure 5.8 compares the average performance of each class of all employed methodologies. The figure shows that the accurate classification of each heartbeat parameter was improved by approximately 5%, which is encouraging progress in this field.

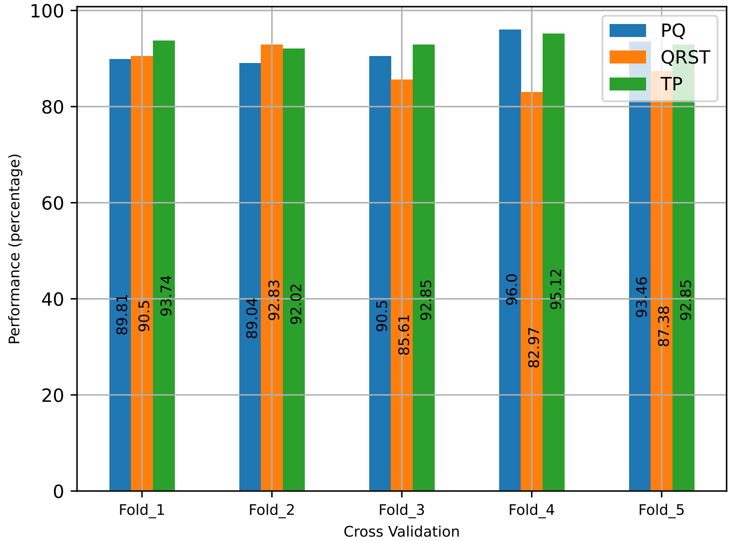

Figure 5.7. Performance of the model when STFT and entropy were added

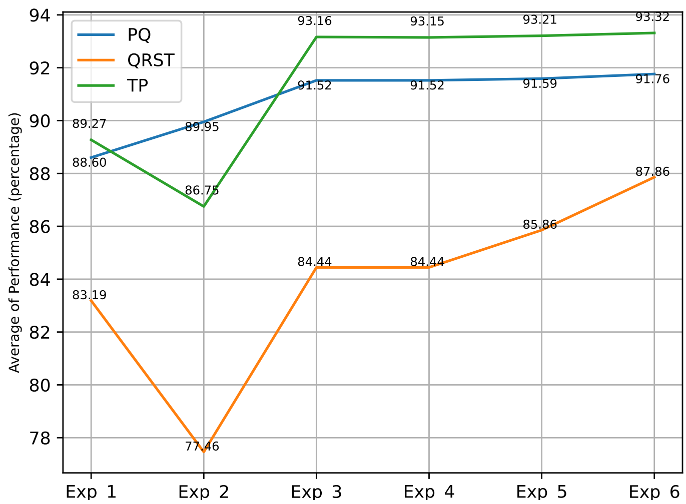

Figure 5.8. The average of the Performance of all methods

5.5 Discussion

Accurately detecting critical parameters of ECG signals is of utmost importance as the consequences of any errors could potentially jeopardize the life and well-being of the patients [69]. Any inaccuracies or errors in ECG signal detection can have profound implications and may lead to adverse health outcomes for patients [71]. Moreover, accurately detecting ECG parameters is a crucial initial step in predicting patients' future health issues [10]. By correctly identifying the ECG parameters, healthcare professionals can determine the current cardiac condition of patients and predict potential health issues that may arise in the future. This information can help healthcare providers develop individualized treatment plans for patients, as well as provide early intervention and preventive care to improve patient outcomes [69]. Therefore, it is essential that healthcare providers have the necessary tools and expertise to accurately detect ECG parameters and interpret the results to provide the best possible care for their patients. Therefore, it is imperative to establish a robust and dependable classification method for classifying the parameters of ECG signals with a high accuracy and low error rate.

In this paper, we classified the ECG signal in order to detect the heart rhythm parameters. we worked on classifying the ECG signal into three main classifications, including 'PQ', 'QRST', and 'TP', through six different methodologies. we applied all proposed methodologies on the two channels PhysioNet's QT public database to test our models. In the first step, we preprocessed the mentioned public dataset by filtering the noises, removing probable gaps in signals, and normalizing them. Then, we feed the DL models with the 3-dimensional arrays. In the current paper, three DL models proposed for detecting ECG parameters, moreover by analyzing the signal at hand STFT and approximate entropy, were used as the additional variables for improving the performance of the model. By employing the suggested methods, the performance of the classification task improved up to 92%. Moreover, all results were generated using a random seed; thus, they are re-generalizable.

The present study aimed to investigate the impact of duplicating layers on the classification performance of ECG signals. Furthermore, the significance of analyzing ECG signals and adding more parameters to the current public dataset was explored. The findings demonstrate that heart rhythm parameters can be accurately classified, attaining a high-reliability level and a low error rate. These results offer valuable insights and provide a solid basis for further research in the field of ECG signal processing toward healthcare 4.0.

Notwithstanding the classification potential of the described methodologies and the opportunity to unveil the role of duplicating DL model layers for analyzing ECG signals and introducing novel parameters to the dataset, it is imperative to be cautious in light of certain considerations. It should be noted that while proposed methodologies may slightly improve the performance of the classification task, they may also increase the time required to analyze and provide results. Therefore, it is imperative to consider the trade-off between performance and time when considering these approaches. As a matter of fact, the results obtained so far are satisfactory. However, providing additional samples and recourse to the algorithms can lead to more precise results and interpretations. It will also help in finding solutions to achieve more accurate outcomes.

5.6 Conclusion

In conclusion, in this chapter, we presented six different approaches based on DL to identify and understand how the ECG signals can be classified with high accuracy and low error rate. The results of the experiments revealed that a significant improvement in accuracy can be achieved by duplicating the layer of the model and extracting values from the signal. This approach presents a promising opportunity to enhance the performance and prove the potential of AI to aid the decision of clinicians to provide more wise services.

The proposed methods were applied to PhysioNet's QT public database. We worked on categorizing ECG signal parameters into three main groups, namely 'PQ', 'QRST', and 'TP', and We obtained accuracy up to 92%. To achieve the desired outcome, we employed a model with a duplicated layer while adjusting several parameters. Additionally, we incorporated approximate entropy and STFT as supplementary features. The use of these features facilitated the extraction of key insights, providing a more comprehensive understanding of the signal.

Although, by utilizing the mentioned sophisticated methods, we were able to achieve optimal results with a high degree of accuracy, future studies on various datasets and powerful computational recourses should confirm our methodology and findings.

Chapter 6

Conclusion

Healthcare 4.0 covers a multitude of domains that are essential for providing exceptional healthcare services to patients and their families. Employing AI-based methods, healthcare providers should confidently make accurate and timely decisions to improve patient outcomes. In this study, our focus was to examine various domains to determine the potential of AI to support clinicians on the path toward Healthcare 4.0. Our research was conducted with rigorous attention to detail, using standardized methods of data collection and analysis. We utilized a genuine database for the covered domains to gain a practical viewpoint on the proposed topic. This approach has enabled us to acquire a practical perspective, which is essential to comprehend the details and complexities of the topic. The study evaluated the effectiveness of AI techniques in four different clinical areas from various perspectives.

First, we focused on prostate cancer. We investigated the effectiveness of ML methods for predicting the patient quality of life after surgery. We concentrated on sexual function as an important indicators of quality of life of these subjects. The model was able to predict target values with a high level of accuracy and with a low rate of errors by using relevant patient data. This was a collaboration work with the hospital of the University of Naples, Federico II. In this work, we worked on knowing critical predictors, and we also achieved the prediction with up to 74% accuracy in the case of binary classification. The findings suggest that ML models can effectively predict sexual function and provide valuable insights into the factors that contribute to sexual health after prostate surgery.

We then considered two different areas in the psychiatric disease field. Firstly, we developed a methodology for understanding the severity of the lockdown situations, such as the COVID-19 pandemic, on psychiatric symptoms. This study was conducted on a small group of participants coming from a real world study. Results show that depression, anxiety, and obsessive-compulsive symptoms during lockdown can be predicted with up to 92% accuracy based on demographic and clinical characteristics collected prior to the pandemic.

The second health-related disease we considered is depression. We studied how to predict the depression level after the innovative tDCS therapy. The AI-based methodology was able to identify important predictors and provide insights into the relationship between depression and demographic and clinical characteristics. Moreover, the study showed the ability of the used method to predict the depression level, measured through the HDRS, with an accuracy of up to 63%. This result was obtained on a small real-world dataset.

Diabetes is another important clinical field that we studied. We considered two different aspects of this field. First, we evaluated the role of key enabling technology in diabetic management by conducting a systematic literature review. The result of this work is a deep analysis of the body of knowledge in the scientific literature as well as in technical documents related to how key enabling technologies are used in the field of diabetes management. In the same field, we focused on predicting glycemia events in diabetes. We worked in collaboration with a life support technology group from The Technical University of Madrid and San Carlos Clinical Hospital in Madrid, Spain. By employing ML methods, not only the correlation of different predictors with the target value and each other is recognized, but also the glycemia event of the diabetics are predicted with 95% accuracy.

Finally, we concentrated on hearth diseases. In particular, we investigated the potential of DL approaches for the analysis of ECG signals. The primary objective was to develop a classification model that could accurately classify ECG signals into relevant heart rate parameters. To this end, six DL approaches based on DL were employed and evaluated to determine their efficacy in reaching the assumed goal. As a result of this analysis, ECG signals are classified into related heart rate parameters with an accuracy of up to 93%.

The mixed-domain research project conducted in this work highlights the potential of employing AI methods in the healthcare sector towards achieving the healthcare 4.0 paradigm. The research findings suggest that using AI in this field can enhance patient outcomes and healthcare quality. The implementation of AI can effectively solve various issues such as medical errors, diagnostic inaccuracy, poor resource utilization, and patient safety. The results of this research indicate that AI can have a transformative impact on healthcare by enhancing the efficiency and accuracy of diagnosis and treatment, leading to a better quality of life for patients.

These results underscore the importance of continued research and development in this area and highlight the necessity of effective collaboration between healthcare professionals, researchers, and technology experts, in particular in terms of collecting well-organized large datasets.

6.1 Limitations and future works

The developed methodologies showcase significant potential and hold promise in unraveling unknown relationships between clinically relevant variables in current research. However, it is imperative to proceed cautiously in interpreting the achieved results due to certain limitations. Therefore, the results should be analyzed with meticulous attention to detail while considering the limitations. First, the results of the present study were based on a relatively small sample size compared with real-world data size, and both clinical and non-clinical samples are probably not representative of the corresponding populations. In some datasets, we did not include healthy subjects. Moreover, in some cases, we encountered a substantial amount of missing data within the dataset. As a result, we either removed certain cases from the dataset or substitute missing values with a fixed number. However, this approach may increase the risk of bias in our analysis. In addition, when there is a limited amount of training data available, AI-based methods might encounter overfitting issues. This happens when the model becomes too focused on the training data and fails to generalize well when faced with test data. To address this, we opted to utilize simple techniques that are better suited to offer improved performance when working with small data samples. Another obstacle we faced was having unbalanced data; we had to adjust the threshold for some classes based on input from clinicians and algorithm performance. The future work of this research study can be improved by using more optimization techniques and more complex classification algorithms. Additionally, training a larger dataset that not only contains more cases but also covers more characters of the participants would be an important extension of this work.

Bibliography

[1] https://www.r-project.org/.

[2] Centers for disease control and prevention. https://data.cdc.gov/.

[3] Organisation for economic co-operation and development. https://www.oecd.org/spain/.

[4] L. Abhishek. Optical character recognition using ensemble of svm, mlp and extra trees classifier. In *2020 International Conference for Emerging Technology (INCET)*, pages 1–4, 2020.

[5] Hedayat Abrishami. Supervised ecg interval segmentation using lstm neural network. 08 2018.

[6] Nnenaya Agochukwu-Mmonu, Adharsh Murali, Daniela Wittmann, Brian Denton, Rodney L Dunn, James Montie, James Peabody, David Miller, Karandeep Singh, et al. Development and validation of dynamic multivariate prediction models of sexual function recovery in patients with prostate cancer undergoing radical prostatectomy: Results from the music statewide collaborative. *European Urology Open Science*, 40:1–8, 2022.

[7] Mahsa Akhbari, Mohammad B. Shamsollahi, Omid Sayadi, Antonis A. Armoundas, and Christian Jutten. Ecg segmentation and fiducial point extraction using multi hidden markov model. *Computers in Biology and Medicine*, 79:21–29, 2016.

[8] Alejandro Albizu, Ruogu Fang, Aprinda Indahlastari, Andrew O'Shea, Skylar E Stolte, Kyle B See, Emanuel M Boutzoukas, Jessica N Kraft, Nicole R Nissim, and Adam J Woods. Machine learning and individual variability in electric field characteristics predict tdcs treatment response. *Brain stimulation*, 13(6):1753–1764, 2020.

[9] Ganjar Alfian, Muhammad Syafrudin, Muhammad Fazal Ijaz, M Alex Syaekhoni, Norma Latif Fitriyani, and Jongtae Rhee. A personalized healthcare monitoring system for diabetic patients by utilizing ble-based sensors and real-time data processing. *Sensors*, 18(7):2183, 2018.

[10] F Alonso-Atienza, E Morgado, L Fernández-Martínez, A García-Alberola, and JL Rojo-Álvarez. Combination of ecg parameters with support vector machines for the detection of life-threatening arrhythmias. In *2012 Computing in Cardiology*, pages 385–388, 2012.

[11] R.V. Andreao, B. Dorizzi, and J. Boudy. Ecg signal analysis through hidden markov models. *IEEE Transactions on Biomedical Engineering*, 53(8):1541–1549, 2006.

[12] Mohammad Reza Askari, Mahmoud Abdel-Latif, Mudassir Rashid, Mert Sevil, and Ali Cinar. Detection and classification of unannounced physical activities and acute psychological stress events for interventions in diabetes treatment. *Algorithms*, 15(10):352, 2022.

[13] Javier Aspuru, Alberto Ochoa-Brust, Ramón A Félix, Walter Mata-López, Luis J Mena, Rodolfo Ostos, and Rafael Martínez-Peláez. Segmentation of the ecg signal by means of a linear regression algorithm. *Sensors*, 19(4):775, 2019.

[14] Javier Aspuru, Alberto Ochoa-Brust, Ramón A. Félix, Walter Mata-López, Luis J. Mena, Rodolfo Ostos, and Rafael Martínez-Peláez. Segmentation of the ecg signal by means of a linear regression algorithm. *Sensors*, 19(4), 2019.

[15] Jahongir Azimjonov and Taehong Kim. Stochastic gradient descent classifier-based lightweight intrusion detection systems using the efficient feature subsets of datasets. *Expert Systems with Applications*, 237:121493, 2024.

[16] Pritika Bahad and Preeti Saxena. Study of adaboost and gradient boosting algorithms for predictive analytics. In *International Conference on Intelligent Computing and Smart Communication 2019: Proceedings of ICSC 2019*, pages 235–244. Springer, 2020.

[17] Junaid Bajwa, Usman Munir, Aditya Nori, and Bryan Williams. Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*, 8(2):e188, 2021.

[18] Sutapa Basu, Swapna Verma, and Edimansyah Abdin. The evolution of depression in first-episode psychosis: A naturalistic database study. *Clinical Psychology & Psychotherapy*, 27(4):611–620, 2020.

[19] Felix D Beacher, Lilianne R Mujica-Parodi, Shreyash Gupta, and Leonardo A Ancora. Machine learning predicts outcomes of phase iii clinical trials for prostate cancer. *Algorithms*, 14(5):147, 2021.

[20] AT Beck, RA Steer, GK Brown, et al. Beck depression inventory. san antonio. *TX: The Psychological Corporation*, 1996.

[21] Emelia J Benjamin, Salim S Virani, Clifton W Callaway, Alanna M Chamberlain, Alexander R Chang, Susan Cheng, Stephanie E Chiuve, Mary Cushman, Francesca N Delling, Rajat Deo, et al. Heart disease and stroke statistics—2018 update: a report from the american heart association. *Circulation*, 137(12):e67–e492, 2018.

[22] James RA Benoit, Serdar M Dursun, Russell Greiner, Bo Cao, Matthew RG Brown, Raymond W Lam, and Andrew J Greenshaw. Using machine learning to predict remission in patients with major depressive disorder treated with desvenlafaxine. *The Canadian Journal of Psychiatry*, 67(1):39–47, 2022.

[23] Idoia Beraza and Iñaki Romero. Comparative study of algorithms for ecg segmentation. *Biomedical Signal Processing and Control*, 34:166–173, 2017.

[24] Idoia Beraza and Iñaki Romero. Comparative study of algorithms for ecg segmentation. *Biomedical Signal Processing and Control*, 34:166–173, 04 2017.

[25] Arthur Bertachi, Clara Viñals, Lyvia Biagi, Ivan Contreras, Josep Vehí, Ignacio Conget, and Marga Giménez. Prediction of nocturnal hypoglycemia in adults with type 1 diabetes under multiple daily injections using continuous glucose monitoring and physical activity monitor. *Sensors*, 20(6):1705, 2020.

[26] Dillon Bowen and Lyle Ungar. Generalized shap: Generating multiple types of explanations in machine learning. *arXiv preprint arXiv:2006.07155*, 2020.

[27] Christian John Brakenridge, Agus Salim, Genevieve Nissa Healy, Ruth Grigg, Alison Carver, Kym Rickards, Neville Owen, and David Wayne Dunstan. The associations of covid-19 lockdown restrictions with longer-term activity levels of working adults with type 2 diabetes: Cohort study. *JMIR diabetes*, 7(2):e36181, 2022.

[28] Matthias Braun, Patrik Hummel, Susanne Beck, and Peter Dabrock. Primer on an ethics of ai-based decision support systems in the clinic. *Journal of medical ethics*, 47(12):e3–e3, 2021.

[29] Fraser Cameron, Günter Niemeyer, Karen Gundy-Burlet, and Bruce Buckingham. Statistical hypoglycemia prediction. *Journal of diabetes science and technology*, 2(4):612–621, 2008.

[30] Marzia Cescon, Divya Choudhary, Jordan E Pinsker, Vikash Dadlani, Mei Mei Church, Yogish C Kudva, Francis J Doyle III, and Eyal Dassau. Activity detection and classification from wristband accelerometer data collected on people with type 1 diabetes in free-living conditions. *Computers in biology and medicine*, 135:104633, 2021.

[31] Chelsea Chandler, Peter W Foltz, and Brita Elvevåg. Improving the applicability of ai for psychiatric applications through human-in-the-loop methodologies. *Schizophrenia Bulletin*, 48(5):949–957, 2022.

[32] Bahzad Charbuty and Adnan Abdulazeez. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01):20–28, 2021.

[33] N Cheffi, O Chakroun-Walha, R Sellami, R Ouali, D Mnif, F Guermazi, F Issaoui, M Lajmi, B Benamar, J Damak, et al. Validation of the hamilton depression rating scale (hdrs) in the tunisian dialect. *Public Health*, 202:100–105, 2022.

[34] Rung-Ching Chen, Christine Dewi, Su-Wen Huang, and Rezzy Eko Caraka. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7:1–26, 2020.

[35] European Commission and Eurostat. *HEDIC – Health expenditures by diseases and conditions – 2016 edition*. Publications Office, 2016.

[36] Andrew JM Cooper, Soren Brage, Ulf Ekelund, Nicholas J Wareham, Simon J Griffin, and Rebecca K Simmons. Association between objectively assessed sedentary time and physical activity with metabolic risk factors among people with recently diagnosed type 2 diabetes. *Diabetologia*, 57:73–82, 2014.

[37] Daniel J Cox, Linda Gonder-Frederick, Lee Ritterband, William Clarke, and Boris P Kovatchev. Prediction of severe hypoglycemia. *Diabetes Care*, 30(6):1370–1373, 2007.

[38] Anna Czmil, Sylwester Czmil, and Damian Mazur. A method to detect type 1 diabetes based on physical activity measurements using a mobile device. *Applied Sciences*, 9(12):2555, 2019.

[39] Darpit Dave, Daniel J DeSalvo, Balakrishna Haridas, Siripoom McKay, Akhil Shenoy, Chester J Koh, Mark Lawley, and Madhav Erraguntla.

Feature-based machine learning model for real-time hypoglycemia prediction. *Journal of Diabetes Science and Technology*, 15(4):842–855, 2021.

[40] Lehel Dénes-Fazakas, Máté Siket, László Szilágyi, Levente Kovács, and György Eigner. Detection of physical activity using machine learning methods based on continuous blood glucose monitoring and heart rate signals. *Sensors*, 22(21):8568, 2022.

[41] Najmeddine Dhieb, Hakim Ghazzai, Hichem Besbes, and Yehia Massoud. Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, pages 1–5, 2019.

[42] Katrin Druschky and Achim Druschky. Mobile biofeedback of heart rate variability in patients with diabetic polyneuropathy: a preliminary study. *Clinical physiology and functional imaging*, 35(5):332–337, 2015.

[43] Dominic B. Dwyer, Peter Falkai, and Nikolaos Koutsouleris. Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, 14(1):91–118, 2018. PMID: 29401044.

[44] Giordano D'urso, Alfonso Magliacano, Sayna Rotbei, Felice Iasevoli, Andrea de Bartolomeis, and Alessio Botta. Predicting the severity of lockdown-induced psychiatric symptoms with machine learning. *Diagnostics*, 12(4):957, 2022.

[45] Jane L Eisen, Katharine A Phillips, Lee Baer, Douglas A Beer, Katherine D Atala, and Steven A Rasmussen. The brown assessment of beliefs scale: reliability and validity. *American Journal of Psychiatry*, 155(1):102–108, 1998.

[46] Tarik Elhadd, Raghvendra Mall, Mohammed Bashir, Joao Palotti, Luis Fernandez-Luque, Faisal Farooq, Dabia Al Mohanadi, Zainab Dabbous, Rayaz A Malik, Abdul Badi Abou-Samra, et al. Artificial intelligence (ai) based machine learning models predict glucose variability and hypoglycaemia risk in patients with type 2 diabetes on a multiple drug regimen who fast during ramadan (the profast–it ramadan study). *Diabetes Research and Clinical Practice*, 169:108388, 2020.

[47] Samer Ellahham. Artificial intelligence: the future for diabetes care. *The American journal of medicine*, 133(8):895–900, 2020.

[48] Marc Fakhoury. Artificial intelligence in psychiatry. *Frontiers in Psychiatry: Artificial Intelligence, Precision Medicine, and Other Paradigm Shifts*, pages 119–125, 2019.

[49] Andrea Fiorillo, Gaia Sampogna, Vincenzo Giallonardo, Valeria Del Vecchio, Mario Luciano, Umberto Albert, Claudia Carmassi, Giuseppe Carrà, Francesca Cirulli, Bernardo Dell'Osso, et al. Effects of the lockdown on the mental health of the general population during the covid-19 pandemic in italy: Results from the comet collaborative network. *European Psychiatry*, 63(1), 2020.

[50] Barbara Fletcher, Meg Gulanick, and Cindy Lamendola. Risk factors for type 2 diabetes mellitus. *Journal of Cardiovascular Nursing*, 16(2):17–23, 2002.

[51] Kazuya Fujihara and Hirohito Sone. Machine learning approach to drug treatment strategy for diabetes care. *Diabetes & Metabolism Journal*, 47(3):325–332, 2023.

[52] Kings Fund et al. The health care workforce in england: Make or break. *Kings Fund.[Google Scholar]*, 2018.

[53] Jordan J Gal, Zoey Li, Steven M Willi, and Michael C Riddell. Association between high levels of physical activity and improved glucose control on active days in youth with type 1 diabetes. *Pediatric Diabetes*, 23(7):1057–1063, 2022.

[54] Rishab Gargeya and Theodore Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017.

[55] Eleni I Georga, José C Príncipe, and Dimitrios I Fotiadis. Short-term prediction of glucose in type 1 diabetes using kernel adaptive filters. *Medical & biological engineering & computing*, 57:27–46, 2019.

[56] Nicole Geschwind, Nancy A Nicolson, Frenk Peeters, Jim van Os, Daniela Barge-Schaapveld, and Marieke Wichers. Early improvement in positive rather than negative emotion predicts remission from depression after pharmacotherapy. *European Neuropsychopharmacology*, 21(3):241–247, 2011.

[57] Asma Ghandeharioun, Szymon Fedor, Lisa Sangermano, Dawn Ionescu, Jonathan Alpert, Chelsea Dale, David Sontag, and Rosalind Picard. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 325–332, 2017.

[58] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

[59] Osea Giuntella, Kelly Hyde, Silvia Saccardo, and Sally Sadoff. Lifestyle and mental health disruptions during covid-19. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[60] Mithat Gönen et al. Receiver operating characteristic (roc) curves. *SAS Users Group International (SUGI)*, 31:210–231, 2006.

[61] Wayne K Goodman, Lawrence H Price, Steven A Rasmussen, Carolyn Mazure, Roberta L Fleischmann, Candy L Hill, George R Heninger, and Dennis S Charney. The yale-brown obsessive compulsive scale: I. development, use, and reliability. *Archives of general psychiatry*, 46(11):1006–1011, 1989.

[62] Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

[63] Michela Gravina, Lorenzo Spirito, Giuseppe Celentano, Marco Capece, Massimiliano Creta, Gianluigi Califano, Claudia Collà Ruvolo, Simone Morra, Massimo Imbriaco, Francesco Di Bello, et al. Machine learning and clinical-radiological characteristics for the classification of prostate cancer in pi-rads 3 lesions. *Diagnostics*, 12(7):1565, 2022.

[64] Adrienne Grzenda, William Speier, Prabha Siddarth, Anurag Pant, Beatrix Krause-Sorio, Katherine Narr, and Helen Lavretsky. Machine learning prediction of treatment outcome in late-life depression. *Frontiers in psychiatry*, 12:738494, 2021.

[65] Andrzej Grzybowski, Piotr Brona, Gilbert Lim, Paisan Ruamviboonsuk, Gavin SW Tan, Michael Abramoff, and Daniel SW Ting. Artificial intelligence for diabetic retinopathy screening: a review. *Eye*, 34(3):451–460, 2020.

[66] Andrew G Guzick, Abigail Candelari, Andrew D Wiese, Sophie C Schneider, Wayne K Goodman, and Eric A Storch. Obsessive–compulsive disorder during the covid-19 pandemic: a systematic review. *Current psychiatry reports*, 23(11):1–10, 2021.

[67] Ilse JM Hagedoorn, Niala den Braber, Milou M Oosterwijk, Christina M Gant, Gerjan Navis, Miriam MR Vollenbroek-Hutten, Bert-Jan F Van Beijnum, Stephan JL Bakker, and Gozewijn D Laverman. Low physical activity in patients with complicated type 2 diabetes mellitus is associated with low muscle mass and low protein intake. *Journal of Clinical Medicine*, 9(10):3104, 2020.

[68] Samantha Hajna, Yan Kestens, Stella S Daskalopoulou, Lawrence Joseph, Benoit Thierry, Mark Sherman, Luc Trudeau, Rémi Rabasa-Lhoret, Leslie

Meissner, Simon L Bacon, et al. Neighbourhood walkability and home neighbourhood-based physical activity: an observational study of adults with type 2 diabetes. *BMC public health*, 16(1):1–8, 2016.

[69] Muhammad Salman Haleem, Rossana Castaldo, Silvio Marcello Pagliara, Mario Petretta, Marco Salvatore, Monica Franzese, and Leandro Pecchia. Time adaptive ecg driven cardiovascular disease detector. *Biomedical Signal Processing and Control*, 70:102968, 2021.

[70] Muhammad Salman Haleem and Leandro Pecchia. A deep learning based ecg segmentation tool for detection of ecg beat parameters. In *2022 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–4, 2022.

[71] Muhammad Salman Haleem and Leandro Pecchia. A deep learning based ecg segmentation tool for detection of ecg beat parameters. In *2022 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–4. IEEE, 2022.

[72] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[73] Fengyi Hao, Wanqiu Tan, LI Jiang, Ling Zhang, Xinling Zhao, Yiran Zou, Yirong Hu, Xi Luo, Xiaojiang Jiang, Roger S McIntyre, et al. Do psychiatric patients experience more psychiatric symptoms during covid-19 pandemic and lockdown? a case-control study with service and research implications for immunopsychiatry. *Brain, behavior, and immunity*, 87:100–106, 2020.

[74] Hajar Hasannejadasl, Cheryl Roumen, Henk van der Poel, Ben Vanneste, Joep van Roermund, Katja Aben, Petros Kalendralis, Biche Osong, Lambertus Kiemeney, Inge Van Oort, et al. Predicting erectile dysfunction after treatment for localized prostate cancer. *arXiv preprint arXiv:2110.00615*, 2021.

[75] Charlotte Hayes and Andrea Kriska. Role of physical activity in diabetes management and prevention. *Journal of the American Dietetic Association*, 108(4):S19–S23, 2008.

[76] Isabella Helmreich, Stefanie Wagner, Jochem König, Ralf Kohnen, Armin Szegedi, Christoph Hiemke, and André Tadić. Hamilton depression rating subscales to predict antidepressant treatment outcome in the early course of treatment. *Journal of affective disorders*, 175:199–208, 2015.

[77] Joseph Henson, Alex V Rowlands, Emma Baldry, Emer M Brady, Melanie J Davies, Charlotte L Edwardson, Thomas Yates, and Andrew P Hall. Physical behaviors and chronotype in people with type 2 diabetes. *BMJ Open Diabetes Research and Care*, 8(1):e001375, 2020.

[78] Pieter MR Herings, Adam H Dyer, Sean P Kennelly, Sean Reid, Isabelle Killane, Louise McKenna, Nollaig M Bourke, Conor P Woods, Desmond O'Neill, James Gibney, et al. Gait characteristics and cognitive function in middle-aged adults with and without type 2 diabetes mellitus: Data from enbind. *Sensors*, 22(15):5710, 2022.

[79] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[80] Felice Iasevoli, Michele Fornaro, Giordano D'Urso, Diana Galletta, Claudia Casella, Mariano Paternoster, Claudio Buccelli, Andrea de Bartolomeis, COVID-19 in Psychiatry Study Group, et al. Psychological distress in patients with serious mental illness during the covid-19 outbreak and one-month mass quarantine in italy. *Psychological medicine*, 51(6):1054–1056, 2021.

[81] Atsushi Ikeda, Hirokazu Nosato, Yuta Kochi, Takahiro Kojima, Koji Kawai, Hidenori Sakanashi, Masahiro Murakawa, and Hiroyuki Nishiyama. Support system of cystoscopic diagnosis for bladder cancer based on artificial intelligence. *Journal of endourology*, 34(3):352–358, 2020.

[82] Asim Jan, Hongying Meng, Yona Falinie Binti A Gaus, and Fan Zhang. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):668–680, 2017.

[83] Anupama Jha, Meenu Dave, and Supriya Madan. Comparison of binary class and multi-class classifier using different data mining classification techniques. In *Proceedings of International Conference on Advancements in Computing & Management (ICACM)*, 2019.

[84] Yazan Jian, Michel Pasquier, Assim Sagahyroon, and Fadi Aloul. A machine learning approach to predicting diabetes complications. In *Healthcare*, volume 9, page 1712. MDPI, 2021.

[85] Ram D Joshi and Chandra K Dhakal. Predicting type 2 diabetes using logistic regression and machine learning approaches. *International journal of environmental research and public health*, 18(14):7346, 2021.

[86] Ozcan Kayikcioglu, Sinan Bilgin, Goktug Seymenoglu, and Artuner Deveci. State and trait anxiety scores of patients receiving intravitreal injections. *Biomedicine hub*, 2(2):1–5, 2017.

[87] Sarineh Keshishzadeh and Saeid Rashidi. A system of biometric authentication based on ecg signal segmentation. In *2014 22nd Iranian Conference on Electrical Engineering (ICEE)*, pages 1873–1877, 2014.

[88] Ladislav Kesner and Jiří Horáček. Three challenges that the covid-19 pandemic represents for psychiatry. *The British Journal of Psychiatry*, 217(3):475–476, 2020.

[89] Kyoung Jin Kim, Jung-Been Lee, Jimi Choi, Ju Yeon Seo, Ji Won Yeom, Chul-Hyun Cho, Jae Hyun Bae, Sin Gon Kim, Heon-Jeong Lee, and Nam Hoon Kim. Identification of healthy and unhealthy lifestyles by a wearable activity tracker in type 2 diabetes: A machine learning-based analysis. *Endocrinology and Metabolism*, 37(3):547–551, 2022.

[90] Young Jae Kim, Ji Soo Jeon, Seo-Eun Cho, Kwang Gi Kim, and Seung-Gul Kang. Prediction models for obstructive sleep apnea in korean adults using machine learning techniques. *Diagnostics*, 11(4):612, 2021.

[91] Esther T Kok, Joseph McDonnell, Elly A Stolk, Herman J Stoevelaar, Jan JV Busschbach, Triumph Research Group, et al. The valuation of the international prostate symptom score (ipss) for use in economic evaluations. *European urology*, 42(5):491–497, 2002.

[92] Miroslav Kubat, Ivan Bratko, and Ryszard S Michalski. *A review of machine learning methods*. Citeseer, 1998.

[93] P. Laguna, R.G. Mark, A. Goldberg, and G.B. Moody. A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg. In *Computers in Cardiology 1997*, pages 673–676, 1997.

[94] Alejandro Jose Laguna Sanz, José Luis Díez, Marga Giménez, and Jorge Bondia. Enhanced accuracy of continuous glucose monitoring during exercise through physical activity tracking integration. *Sensors*, 19(17):3757, 2019.

[95] Benjamin Lam, Michael Catt, Sophie Cassidy, Jaume Bacardit, Philip Darke, Sam Butterfield, Ossama Alshabrawy, Michael Trenell, Paolo Missier, et al. Using wearable activity trackers to predict type 2 diabetes: machine learning–based cross-sectional study of the uk biobank accelerometer cohort. *JMIR diabetes*, 6(1):e23364, 2021.

[96] Juan Li, Jin Huang, Lanbo Zheng, and Xia Li. Application of artificial intelligence in diabetes education and management: present status and promising prospect. *Frontiers in public health*, 8:173, 2020.

[97] Shen Li and Yong Zhang. Mental healthcare for psychiatric inpatients during the covid-19 epidemic. *General Psychiatry*, 33(2), 2020.

[98] Zhifei Li, Hai Liu, Zhaoli Zhang, Tingting Liu, and Neal N Xiong. Learning knowledge graph embedding with heterogeneous relation attention networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[99] Guang-Di Liu, Yu-Chen Li, Wei Zhang, and Le Zhang. A brief review of artificial intelligence applications and algorithms for psychiatric disorders. *Engineering*, 6(4):462–467, 2020.

[100] Hai Liu, Tingting Liu, Zhaoli Zhang, Arun Kumar Sangaiah, Bing Yang, and YF Li. Arhpe: Asymmetric relation-aware representation learning for head pose estimation in industrial human-machine interaction. *IEEE Transactions on Industrial Informatics*, 2022.

[101] Hai Liu, Chao Zheng, Duantengchuan Li, Xiaoxuan Shen, Ke Lin, Jiazhang Wang, Zhen Zhang, Zhaoli Zhang, and Neal N Xiong. Edmf: Efficient deep matrix factorization with review feature learning for industrial recommender system. *IEEE Transactions on Industrial Informatics*, 2021.

[102] Aboli N. Londhe and Mithilesh Atulkar. Semantic segmentation of ecg waves using hybrid channel-mix convolutional and bidirectional lstm. *Biomedical Signal Processing and Control*, 63:102162, 2021.

[103] Liyuan Luo, Bin Feng, Senjun Yang, Ning Zhang, and Shengliang Qiu. Clinical characteristics of moderate–severe obsessive–compulsive disorder in children and adolescents in china. *Journal of International Medical Research*, 48(5):0300060520922679, 2020.

[104] Nicos Maglaveras, Telemachos Stamkopoulos, Konstantinos Diamantaras, Costas Pappas, and Michael Strintzis. Ecg pattern recognition and classification using non-linear transformations and neural networks: A review. *International Journal of Medical Informatics*, 52(1):191–208, 1998.

[105] Bikesh Maharjan, Juan Li, Jun Kong, and Cui Tao. Alexa, what should i eat?: A personalized virtual nutrition coach for native american diabetes patients using amazon's smart speaker technology. In *2019 IEEE International Conference on E-health Networking, Application & Services (HealthCom)*, pages 1–6. IEEE, 2019.

[106] Aman Malali, Srinidhi Hiriyannaiah, Siddesh G.M., Srinivasa K.G., and Sanjay N.T. Supervised ecg wave segmentation using convolutional lstm. *ICT Express*, 6(3):166–169, 2020.

[107] Nicola Marotta, Alessandro De Sire, Cinzia Marinaro, Lucrezia Moggio, Maria Teresa Inzitari, Ilaria Russo, Anna Tasselli, Teresa Paolucci, Paola Valentino, and Antonio Ammendolia. Efficacy of transcranial direct current stimulation (tdcs) on balance and gait in multiple sclerosis patients: a machine learning approach. *Journal of Clinical Medicine*, 11(12):3505, 2022.

[108] Alessio Martino, Antonello Rizzi, and Fabio Massimo Frattale Mascioli. Supervised approaches for protein function prediction by topological data analysis. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018.

[109] Yuan Meng, Nianhua Yang, Zhilin Qian, and Gaoyu Zhang. What makes an online review more helpful: an interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(3):466–490, 2020.

[110] Karel GM Moons, Douglas G Altman, Johannes B Reitsma, John PA Ioannidis, Petra Macaskill, Ewout W Steyerberg, Andrew J Vickers, David F Ransohoff, and Gary S Collins. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Annals of internal medicine*, 162(1):W1–W73, 2015.

[111] Mário W. L. Moreira, Joel J. P. C. Rodrigues, Valery Korotaev, Jalal Al-Muhtadi, and Neeraj Kumar. A comprehensive review on smart decision support systems for health care. *IEEE Systems Journal*, 13(3):3536–3545, 2019.

[112] MK Moridani, M Abdi Zadeh, and Z Shahiazar Mazraeh. An efficient automated algorithm for distinguishing normal and abnormal ecg signal. *Irbm*, 40(6):332–340, 2019.

[113] Othmar Moser, Marlene Dietrich, Olivia McCarthy, Richard M Bracken, and Max L Eckstein. Bolus insulin dose depends on previous-day race intensity during 5 days of professional road-cycle racing in athletes with type 1 diabetes: A prospective observational study. *Diabetes, Obesity and Metabolism*, 22(10):1714–1721, 2020.

[114] Viktor Moskalenko, Nikolai Zolotykh, and Grigory Osipov. Deep learning for ecg segmentation. In *Advances in Neural Computation, Machine Learning, and Cognitive Research III: Selected Papers from the XXI International Conference on Neuroinformatics, October 7-11, 2019, Dolgoprudny, Moscow Region, Russia*, pages 246–254. Springer, 2020.

[115] Antonio Mucherino, Petraq J Papajorgji, Panos M Pardalos, Antonio Mucherino, Petraq J Papajorgji, and Panos M Pardalos. K-nearest neighbor classification. *Data mining in agriculture*, pages 83–106, 2009.

[116] Andreas C Müller and Sarah Guido. *Introduction to machine learning with Python: a guide for data scientists.* " O'Reilly Media, Inc.", 2016.

[117] Franklin R Muntis, Daria Igudesman, Angelica Cristello Sarteau, Joan Thomas, Nora Arrizon-Ruiz, Julie Hooper, Ananta Addala, Jamie L Crandell, Michael C Riddell, David M Maahs, et al. Relationship between

moderate-to-vigorous physical activity and glycemia among young adults with type 1 diabetes and overweight or obesity: Results from the advancing care for type 1 diabetes and obesity network (act1on) study. *Diabetes Technology & Therapeutics*, 24(12):881–891, 2022.

[118] World Health Organization et al. Working for health and growth: investing in the health workforce. 2016.

[119] Oyeyemi Osho and S Hong. An overview: Stochastic gradient descent classifier, linear discriminant analysis, deep learning and naive bayes classifier approaches to network intrusion detection. *International Journal of Engineering and Technical Research*, 10(4):294–308, 2021.

[120] Basak Ozaslan, Stephen D Patek, and Marc D Breton. Impact of daily physical activity as measured by commonly available wearables on mealtime glucose control in type 1 diabetes. *Diabetes Technology & Therapeutics*, 22(10):742–748, 2020.

[121] Meaghan L O'Donnell, James A Agathos, Olivia Metcalf, Kari Gibson, and Winnie Lau. Adjustment disorder: Current developments and future directions. *International Journal of Environmental Research and Public Health*, 16(14):2537, 2019.

[122] Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *bmj*, 372, 2021.

[123] Ulrich Palm, Alkomiet Hasan, Wolfgang Strube, and Frank Padberg. tdcs for the treatment of depression: a comprehensive review. *European archives of psychiatry and clinical neuroscience*, 266:681–694, 2016.

[124] Xiaoying Pan, Rebecca Levin-Epstein, Jiahao Huang, Dan Ruan, Christopher R King, Amar U Kishan, Michael L Steinberg, and X Sharon Qi. Dosimetric predictors of patient-reported toxicity after prostate stereotactic body radiotherapy: Analysis of full range of the dose–volume histogram using ensemble machine learning. *Radiotherapy and Oncology*, 148:181–188, 2020.

[125] Egon S Pearson. Some notes on sampling tests with two variables. *Biometrika*, pages 337–360, 1929.

[126] Marlon Perera, Rohan Mirchandani, Nathan Papa, Geoff Breemer, Anna Effeindzourou, Lewis Smith, Peter Swindle, and Elliot Smith. Psa-based machine learning model improves prostate cancer risk stratification in a screening population. *World journal of urology*, 39:1897–1902, 2021.

[127] The-Hanh Pham, Vinitha Sree, John Mapes, Sumeet Dua, Oh Shu Lih, Joel EW Koh, Edward J Ciaccio, and U Rajendra Acharya. A novel machine learning framework for automated detection of arrhythmias in ecg segments. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–18, 2021.

[128] K Rajalakshmi, S Chandra Mohan, and S Dhinesh Babu. Decision support system in healthcare industry. *International Journal of computer applications*, 26(9):42–44, 2011.

[129] Mudassir M Rashid, Mohammad Reza Askari, Canyu Chen, Yueqing Liang, Kai Shu, and Ali Cinar. Artificial intelligence algorithms for treatment of diabetes. *Algorithms*, 15(9):299, 2022.

[130] Rebecca E Ready, David Watson, and Lee Anna Clark. Psychiatric patient– and informant-reported personality: Predicting concurrent and future behavior. *Assessment*, 9(4):361–372, 2002.

[131] Ravi Reddy, Navid Resalat, Leah M Wilson, Jessica R Castle, Joseph El Youssef, and Peter G Jacobs. Prediction of hypoglycemia during aerobic exercise in adults with type 1 diabetes. *Journal of Diabetes Science and Technology*, 13(5):919–927, 2019.

[132] Ronny Redlich, Nils Opel, Dominik Grotegerd, Katharina Dohm, Dario Zaremba, Christian Bürger, Sandra Münker, Lisa Mühlmann, Patricia Wahl, Walter Heindel, et al. Prediction of individual response to electroconvulsive therapy via machine learning on structural magnetic resonance imaging data. *JAMA psychiatry*, 73(6):557–564, 2016.

[133] Ignacio Rodríguez-Rodríguez, Ioannis Chatzigiannakis, José-Víctor Rodríguez, Marianna Maranghi, Michele Gentili, and Miguel-Ángel Zamora-Izquierdo. Utility of big data in predicting short-term blood glucose levels in type 1 diabetes mellitus through machine learning techniques. *Sensors*, 19(20):4482, 2019.

[134] Tom Rosenström, Pekka Jylhä, C Robert Cloninger, Mirka Hintsanen, Marko Elovainio, Outi Mantere, Laura Pulkki-Råback, Kirsi Riihimäki, Maria Vuorilehto, Liisa Keltikangas-Järvinen, et al. Temperament and character traits predict future burden of depression. *Journal of affective disorders*, 158:139–147, 2014.

[135] Rodolfo Rossi, Valentina Socci, Dalila Talevi, Sonia Mensi, Cinzia Niolu, Francesca Pacitti, Antinisca Di Marco, Alessandro Rossi, Alberto Siracusano, and Giorgio Di Lorenzo. Covid-19 pandemic and lockdown measures impact on mental health among the general population in italy. *Frontiers in psychiatry*, page 790, 2020.

[136] Sayna Rotbei, Luigi Napolitano, Stefania Zinno, Paolo Verze, and Alessio Botta. Predicting patient sexual function after prostate surgery using machine learning. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE, 2023.

[137] Elizabeth Russell-Minda, Jeffrey Jutai, Mark Speechley, Kaitlin Bradley, Anna Chudyk, and Robert Petrella. Health technologies for monitoring and managing diabetes: a systematic review. *Journal of diabetes science and technology*, 3(6):1460–1471, 2009.

[138] A Sapra and P Bhandari. Diabetes mellitus. 2021 sep 18. *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing*, 2022.

[139] Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, and Munam Ali Shah. Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th international conference on automation and computing (ICAC)*, pages 1–6. IEEE, 2018.

[140] Shamsollahi MB. Sayadi O. A model-based bayesian framework for ecg beat segmentation. *Physiol Meas.*, 30(3):335–52, 03 2009.

[141] Birgitte Schantz Laursen. Sexuality in men after prostate cancer surgery: a qualitative interview study. *Scandinavian Journal of Caring Sciences*, 31(1):120–127, 2017.

[142] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[143] Holger Schwenk and Yoshua Bengio. Training methods for adaptive boosting of neural networks. *Advances in neural information processing systems*, 10, 1997.

[144] Mert Sevil, Mudassir Rashid, Iman Hajizadeh, Minsun Park, Laurie Quinn, and Ali Cinar. Physical activity and psychological stress detection and assessment of their effects on glucose concentration predictions in diabetes management. *IEEE Transactions on Biomedical Engineering*, 68(7):2251–2260, 2021.

[145] Mert Sevil, Mudassir Rashid, Zacharie Maloney, Iman Hajizadeh, Sediqeh Samadi, Mohammad Reza Askari, Nicole Hobbs, Rachel Brandt, Minsun Park, Laurie Quinn, et al. Determining physical activity characteristics from wristband data for use in automated insulin delivery systems. *IEEE sensors journal*, 20(21):12859–12870, 2020.

[146] Milap Shah, Nithesh Naik, Bhaskar K Somani, and BM Zeeshan Hameed. Artificial intelligence (ai) in urology-current use and future directions: An itrue study. *Turkish Journal of Urology*, 46(Suppl 1):S27, 2020.

[147] Lloyd S Shapley et al. A value for n-person games. 1953.

[148] Fatemeh Sharifi, Emad Mohammed, Trafford Crump, and Behrouz H Far. A cluster-based machine learning model for large healthcare data analysis. In *International Conference on Big Data Innovations and Applications*, pages 92–106. Springer, 2019.

[149] Fatemeh Sharifi, Emad Mohammed, Trafford Crump, and Behrouz H Far. Explainable analytics to predict the quality of life in patients with prostate cancer from longitudinal data. *Applied Artificial Intelligence*, 36(1):2055393, 2022.

[150] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[151] Mohammadreza Sheykhmousa, Masoud Mahdianpari, Hamid Ghanbari, Fariba Mohammadimanesh, Pedram Ghamisi, and Saeid Homayouni. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:6308–6325, 2020.

[152] Nora Tabea Sibert, Sebastian Dieng, Alisa Oesterle, Günter Feick, Günther Carl, Thomas Steiner, Jörg Minner, Florian Roghmann, Björn Kaftan, Friedemann Zengerling, et al. Psychometric validation of the german version of the epic-26 questionnaire for patients with localized and locally advanced prostate cancer. *World Journal of Urology*, 39(1):11–25, 2021.

[153] Sandra Śmigiel, Krzysztof Pałczyński, and Damian Ledziński. Ecg signal classification using deep learning techniques based on the ptb-xl dataset. *Entropy*, 23(9):1121, 2021.

[154] Giovanni Sparacino, Francesca Zanderigo, Stefano Corazza, Alberto Maran, Andrea Facchinetti, and Claudio Cobelli. Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series. *IEEE Transactions on Biomedical Engineering*, 54(5):931–937, 2007.

[155] CD Spielberger. State-trait anxiety inventory for adults (stai-ad)[database record]. apa psyctests, 1983.

[156] P Naga Srinivasu, G JayaLakshmi, Rutvij H Jhaveri, and S Phani Praveen. Ambient assistive living for monitoring the physical activity of diabetic adults through body area networks. *Mobile Information Systems*, 2022:1–18, 2022.

[157] V Srivardhan. Adaptive boosting of random forest algorithm for automatic petrophysical interpretation of well logs. *Acta Geodaetica et Geophysica*, 57(3):495–508, 2022.

[158] Eberhard Standl, Kamlesh Khunti, Tina Birgitte Hansen, and Oliver Schnell. The global epidemics of diabetes in the 21st century: Current situation and perspectives. *European journal of preventive cardiology*, 26(2_suppl):7–14, 2019.

[159] Mike Stedman, Mark Lunt, Mark Davies, Mark Livingston, Christopher Duff, Anthony Fryer, Simon George Anderson, Roger Gadsby, Martin Gibson, Gerry Rayman, et al. Cost of hospital treatment of type 1 diabetes (t1dm) and type 2 diabetes (t2dm) compared to the non-diabetes population: a detailed economic evaluation. *BMJ open*, 10(5):e033231, 2020.

[160] Dan J Stein, Daniel LC Costa, Christine Lochner, Euripedes C Miguel, YC Reddy, Roseli G Shavitt, Odile A van den Heuvel, and H Blair Simpson. Obsessive–compulsive disorder. *Nature reviews Disease primers*, 5(1):1–21, 2019.

[161] Bharath Sudharsan, Malinda Peeples, and Mansur Shomali. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. *Journal of diabetes science and technology*, 9(1):86–90, 2014.

[162] Konrad M Szymanski, John T Wei, Rodney L Dunn, and Martin G Sanda. Development and validation of an abbreviated version of the expanded prostate cancer index composite instrument for measuring health-related quality of life among prostate cancer survivors. *Urology*, 76(5):1245–1250, 2010.

[163] Yuuki Tazawa, Kuo-ching Liang, Michitaka Yoshimura, Momoko Kitazawa, Yuriko Kaise, Akihiro Takamiya, Aiko Kishi, Toshiro Horigome, Yasue Mitsukura, Masaru Mimura, et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. *Heliyon*, 6(2):e03274, 2020.

[164] Kamuran Turksoy, Colleen Monforti, Minsun Park, Garett Griffith, Laurie Quinn, and Ali Cinar. Use of wearable sensors and biometric variables in an artificial pancreas system. *Sensors*, 17(3):532, 2017.

[165] Erdenebayar Urtnasan, Eun Yeon Joo, and Kyu Hee Lee. Ai-enabled algorithm for automatic classification of sleep disorders based on single-lead electrocardiogram. *Diagnostics*, 11(11):2054, 2021.

[166] William PTM van Doorn, Yuri D Foreman, Nicolaas C Schaper, Hans HCM Savelberg, Annemarie Koster, Carla JH van der Kallen, Anke Wesselius,

Miranda T Schram, Ronald MA Henry, Pieter C Dagnelie, et al. Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The maastricht study. *PloS one*, 16(6):e0253125, 2021.

[167] Andrey M Vaskovsky and Marina S Chvanova. Designing the neural network for personalization of food products for persons with genetic president of diabetic sugar. In *2019 3rd School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR)*, pages 175–177. IEEE, 2019.

[168] Cuiyan Wang, Riyu Pan, Xiaoyang Wan, Yilin Tan, Linkang Xu, Roger S McIntyre, Faith N Choo, Bach Tran, Roger Ho, Vijay K Sharma, et al. A longitudinal study on the mental health of general population during the covid-19 epidemic in china. *Brain, behavior, and immunity*, 87:40–48, 2020.

[169] Jikuo Wang, Xu Qiao, Changchun Liu, Xinpei Wang, YuanYuan Liu, Lianke Yao, and Huan Zhang. Automated ecg classification using a non-local convolutional block attention module. *Computer Methods and Programs in Biomedicine*, 203:106006, 2021.

[170] Yunqiao Wang, Joanna Bhaskaran, Jitender Sareen, Shay-Lee Bolton, Dan Chateau, and James M Bolton. Clinician prediction of future suicide attempts: a longitudinal study. *The Canadian Journal of Psychiatry*, 61(7):428–432, 2016.

[171] Muhammad Wasimuddin, Khaled Elleithy, Abdel-Shakour Abuzneid, Miad Faezipour, and Omar Abuzaghleh. Stages-based ecg signal analysis from traditional signal processing to machine learning approaches: A survey. *IEEE Access*, 8:177782–177803, 2020.

[172] Janet BW Williams. Standardizing the hamilton depression rating scale: past, present, and future. *European archives of psychiatry and clinical neuroscience*, 251:6–12, 2001.

[173] Robert F Wolff, Karel GM Moons, Richard D Riley, Penny F Whiting, Marie Westwood, Gary S Collins, Johannes B Reitsma, Jos Kleijnen, Sue Mallett, and PROBAST Group†. Probast: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, 170(1):51–58, 2019.

[174] World Health Organization. Diabetes. World Health Organization, 2023. Accessed on Mar. 1, 2023.

[175] Konstantia Zarkogianni, Konstantinos Mitsis, Eleni Litsa, M-T Arredondo, G Fico, Alessio Fioravanti, and Konstantina S Nikita. Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus

applying sensors for glucose and physical activity monitoring. *Medical & biological engineering & computing*, 53:1333–1343, 2015.

[176] Li Zhang, Mengya Yuan, Zhen An, Xiangmei Zhao, Hui Wu, Haibin Li, Ya Wang, Beibei Sun, Huijun Li, Shibin Ding, et al. Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: A cross-sectional study of chronic diseases in central china. *PloS one*, 15(5):e0233166, 2020.

[177] Stephen X Zhang, Hao Huang, Jizhen Li, Mayra Antonelli-Ponti, Scheila Farias de Paiva, and José Aparecido da Silva. Predictors of depression and anxiety symptoms in brazil during covid-19. *International Journal of Environmental Research and Public Health*, 18(13):7026, 2021.

[178] Hong Zhao, Chunhui Zhao, and Youxian Sun. Probabilistic abnormal glycemic event alert for t1dm patients. In *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*, pages 795–800, 2017.

[179] Taiyu Zhu, Chukwuma Uduku, Kezhi Li, Pau Herrero, Nick Oliver, and Pantelis Georgiou. Enhancing self-management in type 1 diabetes with wearables and deep learning. *npj Digital Medicine*, 5(1):78, 2022.

[180] R Abu Zitar. Towards neural network model for insulin/glucose in diabetics. *International Journal of Computing & Information Sciences*, 1(1):25, 2003.

[181] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang. Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, 9:515, 2018.

[182] Xiaonan Zou, Yong Hu, Zhewen Tian, and Kaiyuan Shen. Logistic regression model optimization and case analysis. In *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pages 135–139, 2019.

Author's Publications

- Giordano D'Urso, Alfonso Magliacano, Sayna Rotbei, Felice Iasevoli, Andrea de Bartolomeis, Alessio Botta. Predicting the severity of lockdown-induced psychiatric symptoms with machine learning. In *Diagnostics* 12, no. 4 (2022): Page 957
- Rotbei, Sayna, Luigi Napolitano, Stefania Zinno, Paolo Verze, and Alessio Botta. Predicting Patient Sexual Function After Prostate Surgery Using Machine Learning. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages. 1-6. IEEE, 2023.
- Botta, A., Rotbei, S., Zinno, S. and Ventre, G., 2023. Cyber Security of Robots: a Comprehensive Survey. *Intelligent Systems with Applications* Page.200237.