Instituto Superior
de Engenharia
Politécnico de Coimbra

DEPARTMENT OF INFORMATICS AND SYSTEMS ENGINEERING

Reliability of machine learning models based on data density principles applied to Cardiovascular Risk Assessment

Project Report to fulfill the Master's degree in Informatics Engineering

Specialization in Intelligent Data Analysis

Author

José Miguel Baptista de Sousa

Supervisor

Teresa Raquel Corga Teixeira da Rocha

Co-Supervisor

Simão Pedro Mendes Cruz Reis Paredes

INSTITUTO POLITÉCNICO
DE COIMBRA

INSTITUTO SUPERIOR
DE ENGENHARIA
DE COIMBRA

Coimbra, April 2025

ABSTRACT

Machine learning (ML) models have been increasingly adopted across various fields, often achieving remarkable performances. However, in critical areas such as healthcare, where decisions can have significant consequences, a fundamental concern is the ability to trust the predictions generated by these models. By incorporating reliability or confidence measures into prediction outcomes, security and transparency are enhanced, thus promoting trust in human-AI interaction.

The reliability of a ML model can be assessed either as an overall performance measure, using metrics such as sensitivity and specificity, or as a case-specific evaluation, trough pointwise reliability metrics. In the later approach, pointwise reliability can be evaluated based on density principle, which uses distance measures and data patterns to compare a given instance to the examples in the training set.

This study researched and implemented pointwise reliability metrics to evaluate the predictions made by ML models in the clinical context. Three reliability assessment methods based on the density principle were explored for evaluating ML model prediction outcomes, each assigning a reliability level in the range [0,1] to individual predictions: (1) a data proximity based method, which uses distances between data points; (2) a clustering based method, employing the Fuzzy C-Means clustering algorithm; and (3) a convex hull based method, which exploits the convex hull's geometric properties.

These methods were tested on two datasets. The first is a synthetic dataset devised for the classification of the Body Mass Index. The second consists of real-world data collected in a clinical environment, provided by CHUC – Coimbra Hospital and University Centre, comprising information from 1544 acute coronary syndrome patients.

The results show methods (1) and (3) as effective, assigning higher reliability levels to most correct predictions, and lower reliability levels to incorrect predictions. However, the clustering method (2) was unable to establish a clear relationship between prediction outcomes and reliability, showing to be ineffective for this specific purpose. It is worth noting that these results may vary when applied to different datasets.

Based on the results obtained, this research proposes methods with practical application potential for improving clinical decision support. Such approaches may help prevent fatalities by diseases, and provide healthcare professionals with more reliable diagnosis.

Keywords: Pointwise reliability, Density principle, Machine Learning, Cardiovascular Risk Assessment

RESUMO

Os modelos de *machine learning* (ML) têm sido cada vez mais utilizados em diversas áreas, alcançando frequentemente desempenhos notáveis. No entanto, em áreas críticas como a da saúde, onde as decisões podem ter consequências significativas, uma preocupação fundamental quanto ao seu uso é a capacidade de confiar nas previsões que estes modelos geram. Assim, ao incorporar medidas de confiabilidade nas previsões, melhora-se a segurança e a transparência, tornando a interação com os sistemas de IA mais segura e compreensível.

A confiabilidade dos modelos de ML pode ser avaliada recorrendo a uma medida de desempenho global, utilizando métricas como a sensibilidade e a especificidade, ou através de uma medida de desempenho ponto a ponto (*pointwise*), relativa a cada previsão individual. Nesta última abordagem, a confiabilidade *pointwise* de um modelo de ML pode ser avaliada com base em princípios da densidade, os quais utilizam medidas de distância e características dos dados para comparar uma instância com os exemplos presentes no conjunto de treino.

Este estudo investigou e implementou métricas de confiabilidade *pointwise* para avaliar previsões de modelos de ML em contexto clínico. Foram implementados três métodos baseados no princípio da densidade, os quais atribuem a cada previsão um nível de confiança no intervalo [0,1]: (1) método baseado na proximidade dos dados, utilizando as distâncias entre pontos de dados; (2) método baseado em *clustering*, usando o algoritmo *Fuzzy C-means*; e (3) método baseado no conceito de *convex hull*, recorrendo às propriedades geométricas deste polígono convexo.

Para testar os métodos, foram usados dois conjuntos de dados. O primeiro é um conjunto de dados sintéticos para a classificação do Índice de Massa Corporal. O segundo é um conjunto de dados reais recolhidos em ambiente clínico, disponibilizado pelo CHUC – Centro Hospitalar e Universitário de Coimbra, o qual inclui informações de 1544 pacientes com síndrome coronária aguda.

Os resultados indicam que os métodos (1) e (3) são eficazes, atribuindo níveis de confiança mais elevados à maioria das previsões corretamente classificadas e níveis de confiança mais baixos às previsões incorretamente classificadas. Por sua vez, o método (2) não conseguiu estabelecer uma relação clara entre as previsões e a confiança, mostrando-se ineficaz para o objetivo pretendido. Convém salientar que estes resultados poderão variar caso sejam usados conjuntos de dados diferentes.

Com base nos resultados obtidos, esta investigação propõe métodos com potencial prático para o suporte à decisão clínica. Estas abordagens podem ajudar a prevenir fatalidades e garantir diagnósticos mais seguros aos profissionais de saúde.

Palavras-chave: Confiabilidade *pointwise*, Princípio da densidade, *Machine Learning*, Avaliação do risco cardiovascular

AGRADECIMENTOS

A realização deste trabalho não resulta apenas do meu esforço individual, mas também de todo um conjunto de pessoas que, de forma direta ou indireta, contribuíram para a sua conclusão. A elas expresso o meu profundo agradecimento.

À minha orientadora, Teresa Rocha, e ao meu coorientador, Simão Paredes, agradeço pela ajuda e orientação na preparação deste trabalho, assim como pelo conhecimento e contribuições que melhoraram este trabalho. Ao professor Jorge Henriques, pela disponibilidade e sugestões para o desenvolvimento deste projeto.

Aos meus pais e à minha irmã, por todo o apoio incondicional, paciência e constante incentivo durante todo este processo. Agradeço-vos por acreditarem em mim, e por estarem sempre presentes.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACS	Acute Coronary Syndrome
AI	Artificial Intelligence
AMI	Acute Myocardial Infraction
AUC	Area Under the Curve
BCNN	Bayesian Convolutional Neural Networks
BMI	Body Mass Index
CAGR	Compound Annual Growth Rate
CHD	Coronary Heart Disease
CHUC	Coimbra Hospital and University Centre
CP	Conformal Prediction
CRA	Cardiovascular Risk Assessment
CVD	Cardiovascular disease
DL	Deep Learning
DNN	Deep Neural Network
DXA	Dual-energy X-ray Absorptiometry
FCM	Fuzzy C-Means
FHS	Framingham Heart Study
FN	False Negative
FP	False Positive
FRS	Framingham Risk Score
GRACE	Global Registry of Acute Coronary Events
ICM	Intuitive Certainty Measure
ICP	Inductive Conformal Prediction
ICU	Intensive Care Unit
LR	Lasso Regression
MI	Myocardial Infarction
ML	Machine Learning
NB	Naïve Bayes
NSTEMI	Non ST-elevation Myocardial Infarction

OOD	Out Of Distribution
PCA	Principal Component Analysis
PURSUIT	Platelet Glycoprotein IIb/IIIa
SMOTE	Synthetic Minority Oversampling Technique
STEMI	ST-elevation Myocardial Infarction
SVM	Support Vector Machine
TIMI	Thrombolysis in Myocardial Infarction
TN	True Negative
TP	True Positive
XAI	Explainable Artificial Intelligence

1 INTRODUCTION

The Machine Learning (ML) area is rapidly evolving, and its adoption is increasingly common in various fields. One example is the use of ML in the healthcare industry, which has allowed the diagnosis of heart disease [1], prediction of breast cancer [2] and discover new antibiotics [3], among others. However, the rapid advancement of ML leads to the gradual creation of more complex models, originating concerns about transparency, explainability and trust. This issue is particularly relevant in Deep Learning (DL) methods, which often function as a black-box [4][5]. Therefore, it is increasingly important to establish measures that promote confidence in results originated by the models, mainly in medical environments where model errors such as false negative results, can cost human lives.

One area where the risks are particularly critical is the prognosis of cardiovascular disease (CVD), a leading cause of death globally. CVD is a general term for conditions affecting the heart or blood vessels, and it can also be associated with damage in organs such as the brain, heart, kidneys and eyes. The number of deaths caused by CVD worldwide increased from around 12.1 million to 20.5 million between 1990 and 2021 [6]. The tobacco usage, air pollution and high blood pressure are among the leading risk factors of getting CVD. In 2021, approximately 80% of the CVD-related deaths occurred in low- and middle-income countries [7], where access to healthcare is often limited. It is therefore crucial to identify potential CVD in individuals at an early stage.

The integration of ML into CVD detection enables faster and earlier diagnosis compared to traditional methods. However, such results shall not be blindly trusted but used as a support tool for the disease prognosis, alongside other methods, especially when the confidence in the prediction is low. Although most ML models do not directly provide reliability estimates, it is possible to estimate reliability by analyzing the data space and comparing it to the model's predictions. These reliability assessment methods belong to a relatively recent area of research, which is still underdeveloped.

1.1 Objetives

Most of the current methods do not estimate the reliability of individual instances, focusing exclusively on the overall reliability of the model. The current work intends to address this issue.

Thus, the objective of this work is to research and implement reliability metrics that can effectively evaluate individual predictions made by ML models, using model-independent methods. Enhancing existing ML models by integrating reliability metrics, results in improved and dependable prediction models.

This work focuses on the density principle as the main approach to assess the reliability of ML models. The central hypothesis explored is whether the reliability of a specific instance can be assessed based on its structural similarity and proximity to the training instances used in building the model. The final reliability output shall be a value in the range [0,1].

The practical application of this project focuses on assessing the pointwise reliability of machine learning models in critical applications such as healthcare. Specifically, the current work is applied to evaluate the model's ability to predict the 6-month mortality risk for patients after their first cardiovascular event. This evaluation uses data classified by the GRACE scoring system, which is a widely used tool for predicting the risk of death in patients with acute coronary syndrome.

1.2 Main contributions

The main contributions of the developed work are:

- Overview of existing approaches to ensure trust in ML models, with special focus on the density principle.
- Development of methods to assess the pointwise reliability of models, based on state-of-the-art approaches.
- Application of the developed methods on a healthcare environment with real data, more specifically in the cardiovascular risk assessment of patients.

1.3 Task management

This study comprised five main tasks:

- **Task 1**: State of the art related to reliability assessment methods, with special focus on techniques based on density principle, which involves data distribution and structural similarity between instances. Analysis of use cases, particularly cardiovascular risk assessment.
- **Task 2**: Research and implementation of reliability metrics for ML models.
- **Task 3**: Implementation of the required adjustments for use cases.
- **Task 4**: Validation of the developed reliability metrics.
- **Task 5**: Writing of the final report.

The time distribution of each task is presented in Table 1.1.

Table 1.1 - Timeline of tasks

	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dec	Jan	Feb	Mar
T1																			
T2																			
T3																			
T4																			
T5																			

1.4 Document structure

This document is structured in 7 different chapters.

Chapter 2 introduces some background knowledge on CVD and ML for the integration of the reliability concept.

Chapter 3 presents an overview of reliable ML, including a literature review of approaches made to ensure reliable outcomes. It places a particular focus on methods that align with the density principle.

Chapter 4 outlines the methodology, describing the datasets, ML models and pre-processing techniques used. It also presents the developed methods for assessing prediction pointwise reliability.

Chapter 5 presents the results of the proposed reliability methods by its visualization and quantification.

Chapter 6 provides an analysis of the previous results, discussing possible explanations and limitations of the proposed methods.

Chapter 7 addresses the main conclusions, along with possible directions for future developments.

2 BACKGROUND KNOWLEDGE

This chapter addresses several required concepts for the understanding of this study. It involves the development of metrics able to measure how reliable a ML model output is, solely based on the density principle, in the field of cardiovascular risk assessment (CRA). Therefore, an overview of CVD issues, along with concepts such as reliability, density principle and other related topics to ML are important to be highlighted.

The first sub-section provides an overview of CVD, its epidemiology, risk factors and risk score models. The second sub-section provides a brief background on ML, covering basic concepts that serve as a basis for integrating reliability methods that can increase the trustworthiness of such models, in addition to the main topic of this work, the density principle.

2.1 Cardiovascular Risk Assessment

Cardiovascular diseases are a group of disorders of the heart and blood vessels. It's usually characterized by the build-up of fatty deposits inside the arteries and an increased risk of blood clots, but it can also be associated with damage to arteries in organs. It can refer to a number of conditions, such as heart disease, stroke, arrhythmias, among others. The four main types of CVD are: Coronary heart disease (CHD), strokes, peripheral arterial disease, and aortic disease [8]. Acute Coronary Syndrome (ACS) is the acute event of CHD that requires immediate medical attention, including conditions such as Unstable Angina (UA), non-ST-elevation myocardial infarction (NSTEMI) and ST-elevation myocardial infarction (STEMI) [9].

More than four out of five CVD deaths are due to heart attacks, also known as Acute Myocardial Infraction (AMI) which are a type of ACS, and one third of these deaths occur prematurely in people under 70 years old [10]. Symptoms of heart attack include pain or discomfort in the chest or in the arms, and back, difficulty breathing, nausea, faintness and paleness [11].

Heart disease has been the leading cause of mortality in the United States since 1921. Since 1950, death rates from CVD have declined 60%, however the rates haven´t decreased smoothly over the years, and have recently trended upward [12]. In 2021, CVD accounted for 29% of all deaths globally, with approximately 19.4 million deaths worldwide. In Portugal, this effect was of 27% of all deaths, with approximately 32.6 thousand deaths estimate [13]. Over three quarters of CVD deaths take place in low- and middle-income countries. It's forecasted that by 2030 in low-income countries, CVD will be accountable for more deaths than infectious diseases, maternal and perinatal conditions, and nutritional disorders combined [14]. A compelling reason for the primary focus of deaths happening in middle- and low-

income countries is their habitants having less access to effective health services for early detection and treatment, resulting in more people dying due to late prognosis [11]. Future cardiovascular burden is likely to be exacerbated not just by the aging population, but also by the epidemic of obesity and related cardiovascular risk factors, both of which are increasing in prevalence [15].

2.1.1 Risk Factors

Most CVD can be prevented by addressing behavioral and environmental risk factors. There are many risk factors for the disease such as tobacco usage, unhealthy diet and obesity, physical inactivity, harmful use of alcohol (behavioral risk factors) and air pollution (environmental risk factor) [11]. The Framingham Heart Study (FHS) is a well-known study that helped improve the understanding of cardiovascular risk factors. Many of its findings between 1960-1988 led to the discovery of smoking, high blood pressure and cholesterol level abnormalities, for example, as factors that increase the risk of heart disease [16].

Studies show risk factors operate equally in all populations, suggesting that prevention and treatment are likely to be effective independently of the location [17].

2.1.2 Risk score models

Early detection of CVD is of extreme importance, so that management with counselling and medicines can begin. The risk of developing CVD can be estimated by using risk calculators, that use different information about a person related to risk factors, or by clinical assessment. Assessing the cardiovascular risk (the probability that a patient will have a cardiovascular event in a defined period) is done using prognostic models, and even though the treatment depends on the specific condition, some general recommendations exist. These prognostic models calculate a risk percentage of developing a disease, based on the sum of certain risk factors points. There are risk models used for primary and secondary prevention. Risk models used for primary prevention, such as the Framingham Risk Score (FRS), predict the risk of CVD in asymptomatic patients [18]. In contrast, the Thrombolysis in Myocardial Infarction (TIMI) and Global Registry of Acute Coronary Events (GRACE) are examples of secondary prevention risk models, assessing the risk in patients who already experienced cardiovascular events.

A widely used risk model is the FRS, which is a mathematical model for predicting coronary vascular disease during a 10-year period, fed with known risk factors [16]. Individuals with an FRS of 20% or higher are considered to have high coronary heart disease risk equivalents, while a FRS smaller than 10% is considered at low risk. However, FRS presents some limitations due to its biased cohort.

TIMI is another risk model to CRA, developed for short-term prognosis at 14 days [19]. It covers 7 characteristics as risk factors, including age, recent severe angina, ST

deviation, among others, all with the same magnitude. The TIMI and GRACE scores are the most popular and validated ACS prediction models.

The GRACE risk prediction tool is a simplified nomogram, that estimates the cumulative 6-month risk of death or myocardial infarction (MI) for patients with acute coronary syndrome (ACS) [20]. The GRACE risk tool considers the 8 key risk factors for the prediction of death or MI: age, heart rate, systolic blood pressure, serum creatinine, Killip class, cardiac arrest, cardiac troponins and ST-segment deviation. These characteristics are selected as the strongest predictors based on multivariable analysis, quantifying the importance by the respective hazard ratios. The patient's final risk score is the sum of each risk factor value, resulting in the score interval [2-383]. According to this final risk score, [21] categorized the patients in three groups:

- Low-risk: GRACE score ≤ 108 for NSTEMI and UA; GRACE score ≤ 125 for STEMI diagnosis.
- Intermediate-risk: GRACE score 109−140 for NSTEMI and UA; GRACE score 126–154 for STEMI diagnosis.
- High-risk: GRACE score ≥ 141 for NSTEMI and UA; GRACE score ≥ 155 for STEMI patients.

Several ACS risk prediction tools have been devised, with the GRACE score considered the most robust [22]. A comparison between TIMI, PURSUIT and GRACE risk models found that the GRACE model achieves the best predictive accuracy for the purpose of predicting the risk of death or MI at 1 year after admission [19]. Given its effectiveness and popular use, the GRACE score plays an important role in evaluating cardiovascular risk and is therefore employed in this work.

2.2 Machine Learning

To understand ML, it is first necessary to understand AI. There are many definitions of AI, but it can be defined as the science of creating machines that simulate human learning, problem solving, decision making and autonomy. Its research is estimated to have started in the 1940's [23], with big leaps happening in the late 2000's and early 2010's with the development of DL and deep neural networks (DNN). The AI market size is projected to grow from 214.6 billion $ in 2024 to 1,339.1 billion $ in 2030, at a CAGR (compound annual growth rate) of 35.7% [24]. AI can be considered as an assembly of concepts, one of which subfields is ML. Figure 2.1 depicts how these concepts correlate between themselves.

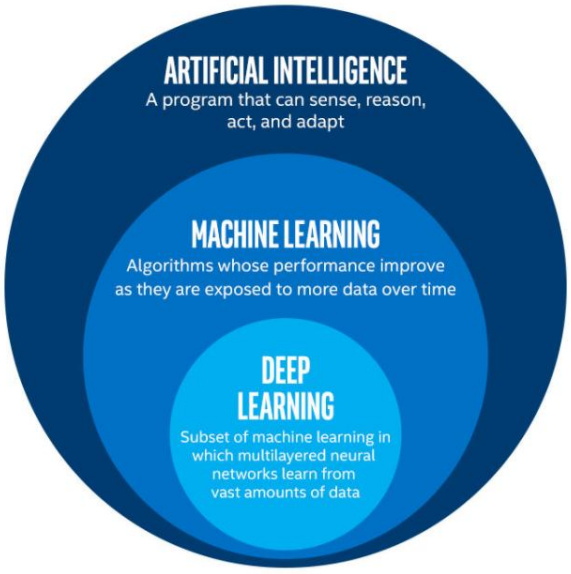


Figure 2.1 - Artificial Intelligence, Machine Learning and Deep Learning [25].

ML is the technique that improves system performance by learning from experience via computational methods [26]. The goal of a ML algorithm is to allow computers to learn automatically without human intervention or assistance and adjust actions accordingly [25]. ML models are organized in two primary categories:

- Supervised learning: Algorithm is trained on labeled data. First generate a model based on processing the dataset and then predicting the label of a new input data point by executing that model [27]. Such examples are classifiers, which map the input space into predefined classes, and regression models, which map the input space into a real-value domain [28].
- Unsupervised learning: Algorithm is given data where labelled examples are not available, using the structural similarity of the data. When new data is introduced, it uses previously learned patterns to recognize the class of the data [29]. Such examples are dimensionality reduction, by reducing the number of features in a dataset with PCA, and clustering, which groups data points in clusters (groups) based on their similarity.

In this work, both supervised and unsupervised ML models play an important role in generating prediction outcomes, which are later evaluated for reliability using model-agnostic methods.

2.3 Interpretability

One of the most popular definitions of interpretability is "the ability to explain or to present in understandable terms to a human" [30]. The more interpretable a ML system is, the easier it is to identify cause-and-effect relationships within the system's inputs and outputs [31]. While ML algorithms perform well, the process by which

they solve a problem is often unclear. This is called the model interpretability problem [27].

One approach to addressing the model interpretability problem is through Explainable AI (XAI), which is a set of methods to allow human comprehension in the model via exploratory data analysis. XAI techniques can be approached in different aspects [32]. Some examples are:

1) **Model being examined:**
    a. Model-specific techniques: Designed purely to interpret models with specific features and capabilities. Deals with inner working of a model to interpret its results.
    b. Model-agnostic techniques: Can be used on any machine learning model. Deals with analyzing features, their relationship with outputs and the data distribution.
2) **Interpretation scope:**
    a. Global interpretation: Involves an overall analysis of a model and its general behavior.
    b. Local interpretation: Involves an analysis of individual predictions and decisions made by the model, to clarify why the model suggested a particular result. When a data point prediction is analyzed, the focus is on the sub-region around that data point.

The work developed in this study intends to concentrate solely on **model-agnostic techniques**, by analyzing the data distribution and only using the prediction outputs from the ML models. Also, the focus is **local interpretation**, as the pointwise reliability will be assessed, which involves the analysis of individual predictions based on spatial data distribution of regions around each individual test case.

2.4 Reliability

Reliability refers to the consistency of repeated measurements of the same event by the same process [33], or the extent to which information can be trusted [34]. In the context of ML, this translates to the ability of a model to consistently produce accurate and stable predictions, or the extent to which it's predictions, the information output, can be trusted. Reliability can be measured at different levels:

- Global reliability: Assessed though qualitative properties of the system related to critical performance indicators, such as accuracy [35].
- Pointwise (individual) reliability: Assessed by estimating the probability that a model's predicted class is equal to the true class for a single instance.

The pointwise reliability concept assesses the model output for each new input, rejecting the output when it's deemed unreliable [36]. This concept is closely related

to Out-Of-Distribution (OOD) detection, because if a test point is far from all the training points, the model should indicate low reliability/reject the output.

Several ML models provide ways to address this concept based on probabilistic distributions, equally in supervised learning as in unsupervised learning. Clustering is an example of unsupervised learning technique, because it uses the similarity properties to group points, and therefore identify data points that fall out the general distribution of known cases [37]. Naïve Bayes (NB) model is an example of supervised learning, computing the probabilities of a data point belonging to each possible label [38], treating features as conditionally independent, which can serve as an indirect measure of reliability for each input.

The trust placed in a ML model should never be absolute, for a black-box nature can never be completely understood, hence its name. This aligns with a fundamental principle from quantum physics: nothing is entirely predictable and there is always uncertainty. The physicist Werner Heisenberg, introduced the Uncertainty Principle, which states that no amount of information can successfully predict the behavior of waves and particles [39]. In other words, the more accurately one property is measured, the less accurately the other property can be known. This concept is carried on to ML in [40], where it's analysis highlights that as neural networks achieve higher accuracy, their vulnerability to adversarial attacks increases.

Uncertainty is the absence of certainty or a condition when the definite outcome or state cannot be established [41]. This means that when a model is uncertain, it won't produce definite outcomes, implying it might be inconsistent. Considering those definitions, it's implicit that reliability and uncertainty are inversely proportional. A reliable model is a consistent model that promotes trust for the user, possessing low uncertainty. Assessing the reliability, indirectly assess the uncertainty. Therefore, the concepts 'reliability' and 'trust' can reflect the same meaning throughout this work, as well the 'reliability assessment' and 'uncertainty assessment' concepts.

There are two main types of uncertainty quantification [42]:

- Aleatoric uncertainty (Data uncertainty): This type of uncertainty arises from inherent noise in the data, such as measurement error or ambiguous annotation, which cannot be reduced by collecting more data.
- Epistemic uncertainty (Model uncertainty): This type of uncertainty arises from a lack of knowledge or information about the underlying model or data distribution or insufficient model structure.
    - Distributional uncertainty: Refers to the uncertainty in the model's predictions when the inputs belong to a different distribution than the training data, i.e., OOD.

This work primarily quantifies epistemic uncertainty, because the uncertainty is calculated due to the lack of knowledge about unseen data distribution, by checking if a test point can be represented in the training distribution. In addition, distributional uncertainty is also used in this work, by the detection of OOD cases.

2.4.1 Density principle

The computation of pointwise reliability can be performed based on two criteria [36]:

- Density principle: Determines whether test case is close to training samples (similar to outlier detection or out-of-distribution samples detection);
- Local fit principle: Determines whether the model was accurate on training samples close to the test case.

While the local fit principle is dependent of the model, the density principle only makes usage of the dataset characteristics to assess reliability. Figure 2.2 depicts a generic application example of the density principle, where the true class labels of the test samples are unknown. Considering the model assigns all test samples to class 0, the reliability levels vary based on their distances to the cluster of class 0 training points. Specifically, the reliability predictions order should be: *Rel(Test sample 1)* > *Rel(Test sample 2)* > *Rel(Test sample 3)*. This is due to the fact that *Test sample 1* is closest to the cluster of class 0 training points, while *Test sample 3* is the farthest. The core idea is that data points near dense regions of training data are more reliable, than data points in isolated regions. Thus, evaluating if a point is far or near training points is forcefully related to the concept of distance between two points.

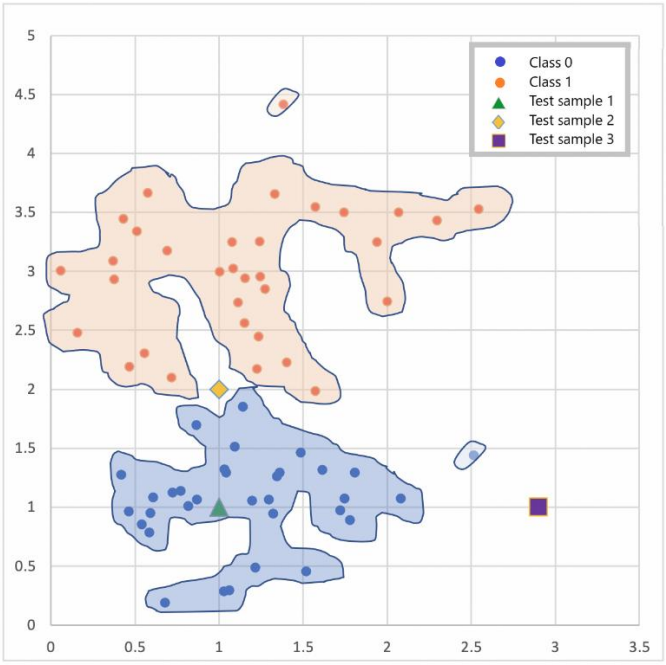


Figure 2.2 – Density Principle. Adapted from [43]

This example is easy to understand because of 2D space distribution. However, the spatial distribution of points becomes harder to visualize and quantify as the number of dimensions' increases. Given a target in high dimensional space, the ratio between

the nearest and farthest neighbor distances is almost 1 for a wide variety of data distributions and distance functions [44], losing meaningfulness from a qualitative perspective in the concept of proximity. Even though computing distance in high dimensions is impractical, dimensionality reductions can be made to achieve compact representations [45]. There are many different distance measurements refined for different types of data.

A. Distance measurements

The Euclidean distance is the most commonly used due to its simplicity. It's mostly used on cases where features have about the same scale and are continuous. The Euclidean distance, d, between two points, x and y, in a *n*-dimensional space, is given by the formula (2.1), where x_k and y_k are respectively the k^{th} attributes of x and y [46].

$$d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2} \tag{2.1}$$

The Gower distance is a measure that can be used to calculate the distance between two entities whose attributes have a mixed of categorical and numerical values. The Gower similarity between two individuals *i* and *j* is given by formula (2.2), where *v* is the number of features, s_{ijk} denotes the contribution provided by the *k*-th variable, and δ_{ijk} is usually 1 or 0 depending if the comparison is valid for the *k*-th variable [47].

$$S_{ij}=\frac{\sum_{k=1}^{v} s_{ijk}\delta_{ijk}}{\sum_{k=1}^{v}\delta_{ijk}} \tag{2.2}$$

The scores of s_{ijk} are assigned differently depending on the types of variables. For quantitative (continuous) variables the formula is given by (2.3), where R_k is the range of *k*. For qualitative (categorical) variables $s_{ijk}=1$ if $x_i=x_j$, else $s_{ijk}=0$. And for dichotomous (binary) variables $s_{ijk}=1$ if $x_i=1$and $x_j=1$, else $s_{ijk}=0$.

$$s_{ij}=1-\frac{|x_i-x_j|}{R_k} \tag{2.3}$$

The formula (2.2) outputs a similarity value in [0,1], where 1 implies the individuals are equal. The distance between the points is computed with formula (2.4).

$$D_{ij}=1-S_{ij} \tag{2.4}$$

There are other distance functions, but this work takes particular importance on these two types of distance: Euclidean distance and Gower distance, due to their relevance to different data type usage.

3 LITERATURE REVIEW

In order to understand the development of reliability in ML, a systematic review of the literature in this domain was carried out. Papers in the subject area of Computer Science, published in the last decade (between 2015 and 2024), were retrieved from Science Direct database. The search queries used were:

- **Search 1**: "((Predictive uncertainty OR Predictive reliability) AND 'machine learning')".
- **Search 2**: "((Predictive uncertainty OR Predictive reliability) AND 'density principle' AND 'machine learning')".

The search 1 identifies all articles that discuss predictive uncertainty or predictive reliability in the context of machine learning, as a global search. Meanwhile, the search 2 works as a more specific search, identifying articles that explore the same topics but also discussing the density principle. The usage of two different searches was chosen so that it's possible to have a general perspective of the development of reliability in ML, while also noticing and comparing with the progress made on this work's specific focus, which is the density principle.

Figure 3.1 represents the annual scientific production of articles in the area of reliability applied to ML. The research interest in this field started with a moderate increase, and became more noticeable around 2019. A possible reason is the advances made on deep learning, such as the emergence of transformers at 2017, which forced a special attention to understanding model predictions. However, the most noticeable jump occurred in the last year (2024) most likely due to the on-going developments and applications of ML and AI for diverse tasks, observing a growing interest in the density principle. The predictive uncertainty and reliability in ML has been growing significantly and, even though the density principle might have started as a niche topic, it also has been growing steadily.

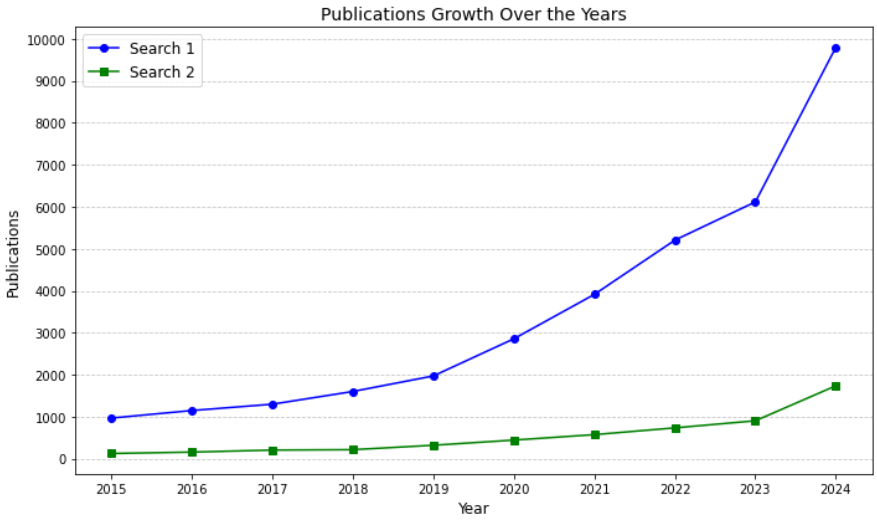


Figure 3.1 – Production of articles evolution in reliable ML.

The literature review to be analyzed was selected for being in English language and open access, and sorted by relevance or by suggestion. Even though this chapter addresses published articles since 2015, various methods to analyze reliability have been developed along the years.

Its origins seems to trace back to the 1990s, where in 1994 [48] uncertainty sampling was used on text categorization tasks. Initially, a fast probabilistic classifier, similar to Naïve Bayes but using logistic regression for scaling, estimates class probabilities for unlabeled data. The instances with probabilities close to 0.5 are considered the most uncertain, and therefore selected for expert labelling. These samples are then used to train another classifier (C4.5 rule induction program), with a loss ratio applied to address class imbalance. The results show that with only an uncertainty sample of 999 instances, C4.5 achieved lower error rates than with random samples of 10,000.

Another example is the confidence quantification, which was addressed as conformal prediction (CP), a statistic framework that enables the calculation of prediction regions with precise confidence levels [49]. CP uses past experience to determine precise levels of confidence in new predictions. It was firstly introduced in 1999 [50] as an underlying algorithm, providing confidence values for its predicted classifications, and also generate a measure of "credibility" which serves as an indicator of how suitable the training data is for classifying the example. In 2007, a modification of the original CP method, called Inductive Conformal Prediction (ICP), was done in order to address the computational efficiency disadvantage of the method, replacing the transductive inference used in the original approach with inductive inference [51].

3.1 Reliability in Healthcare environment

More recently, an example of reliability assessment in healthcare is the use of classifications with a reject option in gene expression data, for ambiguous cases. In this approach, a rejection region is defined in the feature space [52], containing the examples for which classification is ambiguous. The rejection region is determined on a user-defined error rate, and so the algorithm minimizes the rejection region while maintaining the error constraint. This method improves the classifier accuracy, but it's not optimal when the sample size is too low to obtain reliable class density estimations, and when the base classifier lacks discriminatory power. This method's ability to identify and reject ambiguous cases presents an opportunity to reduce wrong diagnosis, when applied to other healthcare scenarios.

Building on the idea that uncertainty estimation plays a crucial role in reducing misdiagnoses, the work in [53] assesses uncertainty using Bayesian Convolutional Neural Networks (BCNN) to improve diagnostic accuracy for COVID-19 detection from chest X-ray images. The experiment used a pre-trained ResNet50V2 model with Dropout layers, where each pass through the network produces a probability

distribution over the predicted classes. Predictive entropy is used as the uncertainty measure, by calculating the entropy of the probability distribution over the classes. Results showed that the estimated uncertainty is higher for erroneous predictions, proving uncertainty estimation can point radiologists for further investigation.

In [54] the risk hip fracture prediction is processed through two stages, using 2 ensemble ML models. For each patient in the test dataset, predictions from all 50 individual logistic regression classifiers from Ensemble 1 were generated, and then calculated the mean and standard deviation of these predictions. The mean symbolizes the final predicted risk, while the standard deviation represents the uncertainty measure. If the standard deviation is ≤ of a threshold, and the mean deviates from 0.5 of a set halfway threshold, then the prediction is considered reliable. In the other hand, if the results don't meet those conditions, the prediction is deemed unreliable and subjected to Ensemble 2, necessitating additional DXA testing to ensure accuracy. The process of this method is seen in Figure 3.2. The results showed most of the patients (≈54%) did not need DXA scans, based on Stage 1 uncertainty assessment, while maintaining high accuracy and AUC values (>80%), reducing diagnostic cost and exposure to radiation.

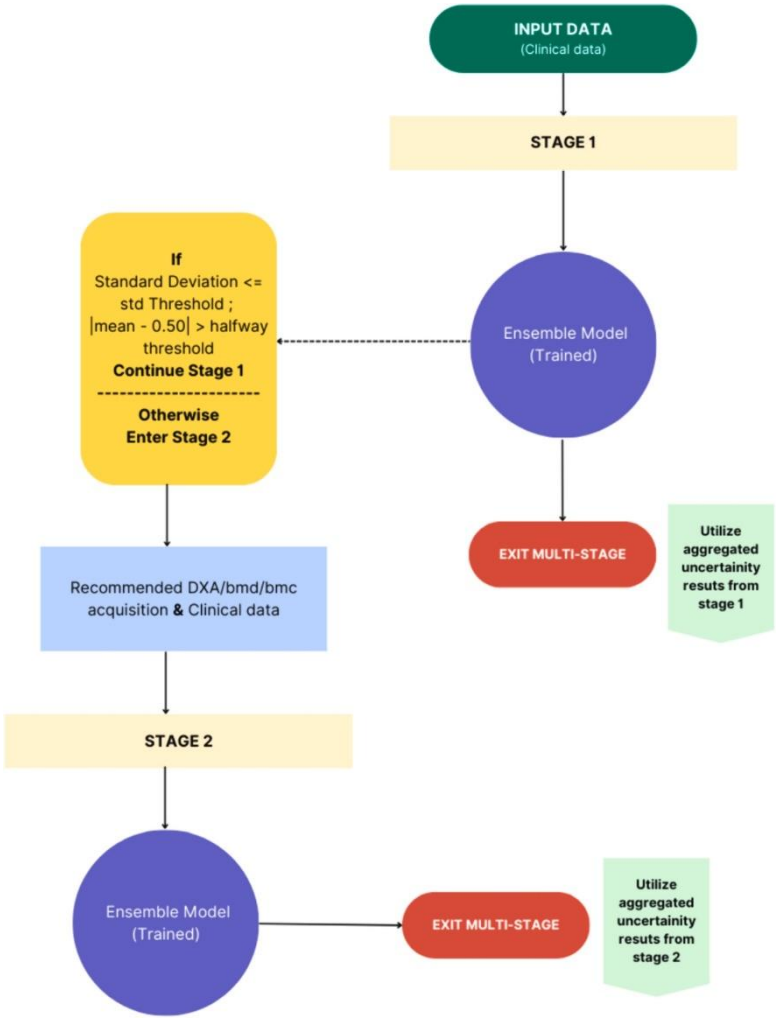


Figure 3.2 – Staged process [54]

The work in [9] uses the same environment as the current study, which is the prediction reliability of the mortality risk after ACS. They do this by creating a set of binary rules based on dichotomized risk factors, separating between positive (death) and negative (survival) classes with a threshold. These rules were trained with ML classifiers that produce a probability output (Logistic Regression and Neural Networks) to predict the probability that each rule is correct for each patient. The reliability estimation is computed by the absolute difference between the means of the predicted acceptances of rules indicating a positive output (death) and rules indicating a negative output (survival), as in (3.1), where p is the number of rules that suggest the patient's death and q is the number of rules that suggest the patient's survival.

$$\text{rel}(x) = \left| \frac{1}{p} \sum_{j=1}^{p} (\text{predicted rule acceptance})_j - \frac{1}{q} \sum_{k=1}^{q} (\text{predicted rule acceptance})_k \right| \quad (3.1)$$

The density principle was firstly introduced in 2019 along the local fit principle, which evaluates the model's accuracy on training samples closest to the test case [36].

The work in [35] applies the local fit principle to a clinical case of in-hospital death predictions. The local fit principle works by selecting the k nearest examples in the training set for each test instance, and calculate the performance of the classifier on these k examples. If the performance metrics are above a predefined threshold, the test instance is classified as "trustful". Both in the simulated dataset as in the real-world clinical dataset, the results obtained from a combination of the density principle followed by the local fit principle, were shown promising with the principles effectively distinguishing reliable from unreliable predictions.

3.2 Density principle

The density principle examines whether a test sample belongs to the distribution in the training set's attribute space. In a previous work by [55], the reliability is computed based on the similarity between the test sample and the training set. This concept is applied to distinguish tumor and normal cells in Acute Myeloid Leukemia patients. To do so, it's necessary to previously detect the border instances for each attribute. Given a binary problem, an instance is considered border of its class for each attribute if it's the nearest (inner border) or the farthest (outer border) to an instance of the opposite class. In the example of Figure 3.3, where each class is represented by a color (red or blue) on a 2D problem:

- For attribute x_1, *A3* and *B1* are inner borders, and *A2* and *B2* are outer borders.
- For attribute x_2, *A3* and *B3* are inner borders, and *A4* and *B1* are outer borders.

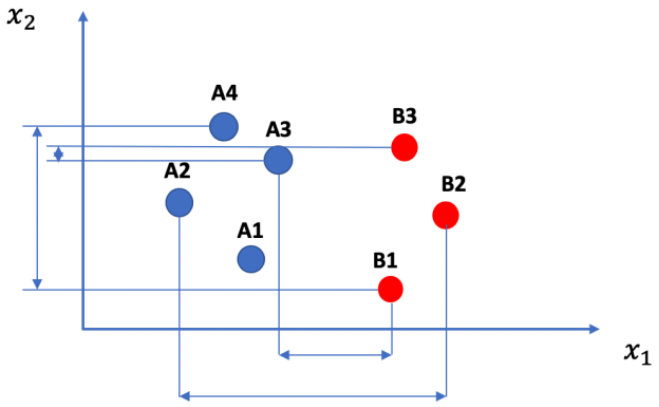


Figure 3.3 – Examples of border instances on two attributes [55]

The similarity is calculated using the equation (3.2), where m is the number of features, and m_x is the quantity of features for which the value of the test sample, x, would be a border instance. If the function is greater than a specified threshold, x is classified as reliable.

$$\text{rel}(x) = 1 - \frac{m_x}{m} \tag{3.2}$$

Results show this model-independent approach is able to partition the validation in different sets, with reliable instances that have higher performances. However, a drawback of this approach is it considers attributes to be independent from each other, something that usually doesn't occur in real world scenarios.

This concept is further explored in [35], where the exact same method is applied and explicitly defined as a density principle based approach. The results showed this principle successfully identified reliable and unreliable samples, with the reliable subset consistently achieving higher accuracy and AUC than the unreliable subset.

In [56] it's proposed an intuitive certainty measure (ICM) which produces an accurate estimation of how certain a ML model is for a specific output, based on errors it made in the past. This measure was applied on synthetic classification problems as well on a real-world high-dimensional dataset. The certainty is measured on the notion of a data point's proximity to other data points in the dataset. If the model's output for a test point (y_x) matches the label of training points (y), then the certainty is higher the closer these points are. On the other hand, if the model's output is different of training points label's, the uncertainty is increased.

The certainty is calculated as the ratio of contributions from training points with $y=y_x$ to the total contributions of all training points. A user defined parameter σ

influences this calculation, by determining the neighborhood size around the test point. The lower σ is, more importance the training points closest to the test case are. To reduce computational costs, ICM only selects the most informative training points instead of using the entire dataset. This selection is done sequentially, based on factors such as the time since a training point was last added, the distance between the point from the current the set, and the number of similar classes examples in the set. Figure 3.4 illustrates how the ICM utilizes the dataset and its parameters to assess the reliability of a test point, on a 2D scenario. By analyzing the proximity of training points to the test point, ICM determines the certainty of the model's prediction.

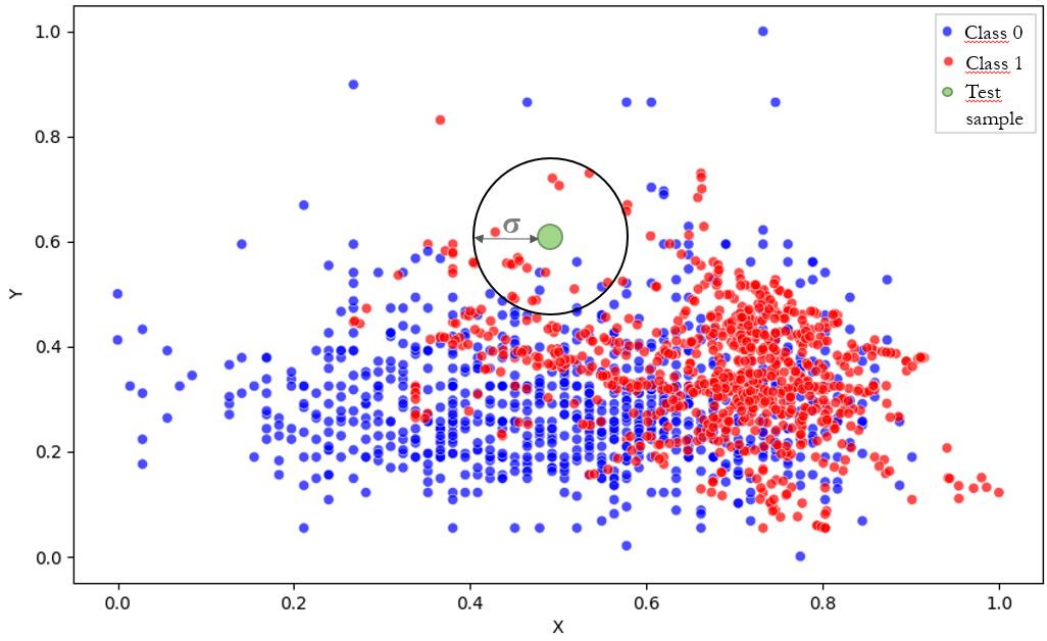


Figure 3.4 - ICM Based Approach

ICM was compared with a baseline model that uses a second ML model, to estimate the probability of correctness. The ICM outperformed the baseline model in all datasets. It achieved higher certainty levels for TP and TN cases, while reducing certainty levels for FP and FN.

The report [43] proposes three trustworthiness metrics in the context of malicious PDF detection, based on the geometric relations between data points and decision boundaries:

(1) **Training Proximity:** A test point (*P*) is deemed trustworthy if it is close to a high density region (cluster) of training points with the same class as the predicted for *P*. Scores range from 0 to 1 (with 1 being the most trustworthy possible outcome) and are computed using HDBSCAN.

(2) **Extrapolation:** identifies *P* trustworthy if it lies within the convex hull formed by the training data, regardless of the class. Score values are unconstrained, but its more trustworthy the closer it is to 0.

(3) **Class Ambiguity**: *P* is considered less trustworthy if it is close to the decision boundary. Produces only non-negative values, but its less trustworthy the closer it is to 0.

These three metrics were applied on the identification of malicious PDFs, using the EvadeML set consisting completely of malicious PDFs. The average result of metric (1) was 0, identifying all the malicious PDFs as outliers. The metric (2) was 0.084, identifying the PDFs as extrapolations, meaning outside of the limit region. And the metric (3) obtained 0.017, meaning all of the data points have similar distances to the decision boundary. All together, the results showed the three metrics are capable of identifying malicious PDFs as untrustworthy.

The local fit principle can be applied together with the density principle to estimate pointwise reliability. In [57] these two principles are employed in a CRA context, to distinguish between reliable and unreliable predictions. The density principle states that a prediction is reliable if there are samples in the training data that are similar to this instance. This component of the reliability assessment is computed as in equation (3.3), where *neighbors(x)* is the number of training points similar to the current instance, assuming a threshold that defines the distance limits within a neighbor is similar to the current instance, and *KMAX* is a maximum value of neighbors considered. The normalization ensures the density result is always on the interval [0,1].

$$\text{densityRel}(x) = \frac{\min(\textit{neighbors}(x), \textit{KMAX})}{\textit{KMAX}} \tag{3.3}$$

4 RESOURCES AND METHODS

This chapter describes the resources and methods used to assess the pointwise reliability. It starts by describing the methodology employed throughout the project, followed by the datasets, ML models and developed methods.

4.1 Methodology

The proposed methodology comprises two phases: i) exploratory phase; ii) real application phase. The Exploratory phase consisted on the search and initial implementation of reliability methods, while the next phase focused on the usage of the best reliability methods previously implemented and less search for new ones. Moreover, another key difference distinguishing the phases is the use of different datasets for each. The Exploratory phase utilized a synthetic dataset, while the Real application phase employed a real-world dataset. Each dataset will be described in detail on the chapter 4.2.

The Real application phase only took place once it was reached a minimum amount of two different reliability methods with promising results from the Exploratory phase. Later, as additional experimentation, some methods that had not transitioned to the following phases due to lack of promising results, were still tested in the Real application phase, in order to determine if the dataset had any influence on the outcomes. Each phase consisted of 5 stages (Figure 4.1):

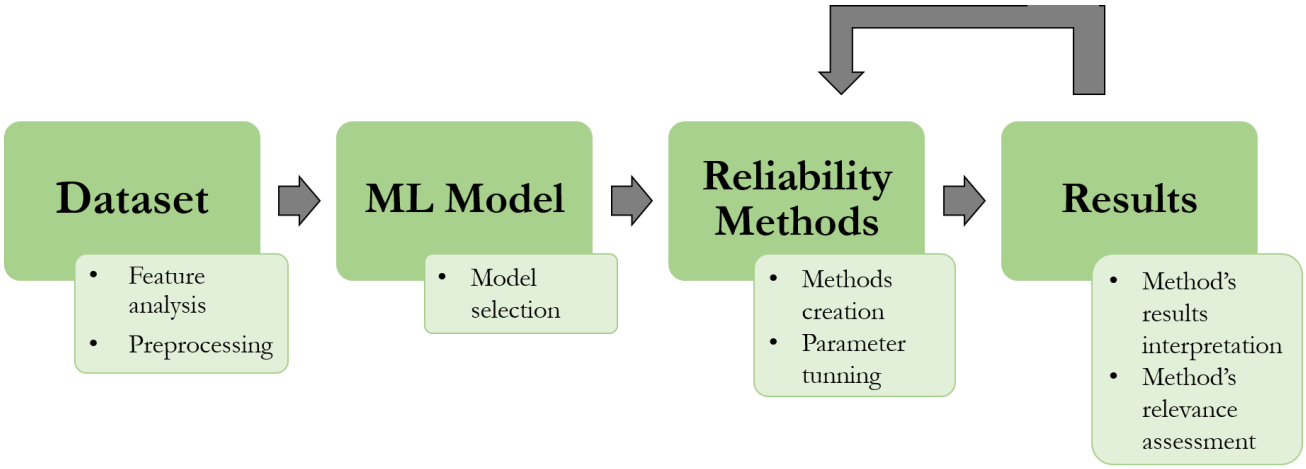


Figure 4.1 – Methodology stages

1. **Dataset:** Introduction of datasets used on the study. Several analyses are required for understanding the data. Identifying correlations and data problems, such as inaccuracies or missing values, are some examples of feature analysis based on the type of problem trying to solve. Preprocessing

techniques, such as cleaning and transforming the data, are steps needed to prepare the data for ML models.

2. **ML Model:** Selection and training of ML models that best suit the problem being attended, chosen for their ability to achieve high evaluation metrics while still maintaining the presence of some incorrect data, to later be analyzed by the methods.
3. **Reliability methods:** Creation or adaptation of methods that evaluate the ML model outputs, producing values in the range [0,1] to quantify the reliability of predictions
4. **Results:** Examination of the reliability methods results using the test set. The results are analyzed to determine the effectiveness and relevance of the methods. Depending on this analysis, methods that are unable to capture a clear relationship between reliability values and prediction outcomes are considered inefficient and therefore discarded. In contrast, methods that show potential may be attempted enhancement by returning to the previous stage.

As a general overview, Figure 4.2 presents the scheme adopted in this work. The pointwise reliability assessment considers a trained ML model, from which will only be used the predicted output of the new instance (model-agnostic and local interpretation) for the reliability metric. With this predicted output, the reliability method compares the new instance with other instances in the dataset, taking into consideration their labels. The final reliability is expressed on a scale from 0 to 1, where 1 indicates a fully reliable prediction.

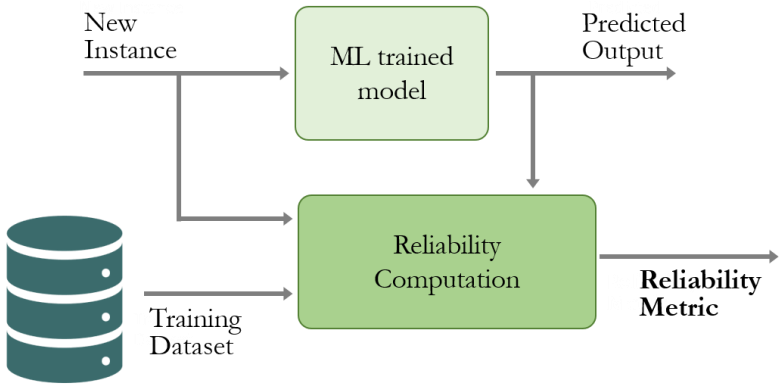


Figure 4.2 - Methodology scheme

4.2 Datasets

This work utilizes two datasets, each corresponding to a distinctive phase of the work development. The Body Mass Index dataset was used in the initial phase for methods development and testing. In the later stage, the Cardiovascular Risk Assessment dataset was used to apply the methods in a more practical and relevant

healthcare scenario. Using both datasets provides a broader validation of the methods performance across different contexts.

4.2.1 Body Mass Index

The Body Mass Index (BMI) dataset is a synthetic dataset delivered in an excel sheet, made to represent the BMI of different people, based on body characteristics. Each BMI value is categorized in 4 classes scaling gradually from underweight to obesity: 0, 1, 2 and 3. Figure 4.3 indicates the BMI values range for each categorized class. As observable, the data follows the standard BMI classification commonly used in the literature [58], where:

- Class 0 with values [15.26–18.49], represents underweight (<18.5);
- Class 1 with values [18.56-24.99] represents normal weight (18.5-24.9);
- Class 2 with values [25.00-29.99] represents overweight (25.0-29.9);
- Class 3 with values [30.01-57.18] represents obesity (>30.0).

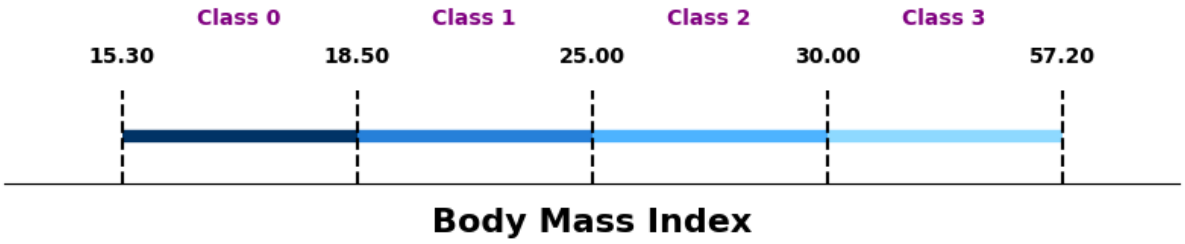


Figure 4.3 – BMI values range for each categorized class

The dataset contains information about 4464 records, and encompass statistics across 5 columns. Table 4.1 shows the characteristics of each of the 4 features. The BMI class is categorized based on the BMI feature, which is calculated based on the remaining 3 features: chest circumference, hip circumference and waist circumference.

Table 4.1 – BMI Dataset feature description.

Feature	Description	Data type	Unique values	Missing data
BMI	Individual body mass index	Numeric (float)	4386	7
Chest circumference	Individual chest perimeter	Numeric (float)	572	7
Hip circumference	Individual hip perimeter	Numeric (float)	522	17
Waist circumference preferred	Individual waist perimeter	Numeric (float)	634	13

Overall, a very low percentage of missing data was found to be present in the dataset, with only 0.16% at BMIclass, BMI and chest circumference features, about 0.38% at hip circumference feature and around 0.29% at waist circumference feature. Altogether there are 29 rows with missing data, with 5 records missing all features and 2 missing both BMIclass and BMI. Since the quantity of missing data is so low and it covers multiple features, those records were removed from the dataset. The Figure 4.4 show the quantity of records per BMI class after the missing data cleanse.

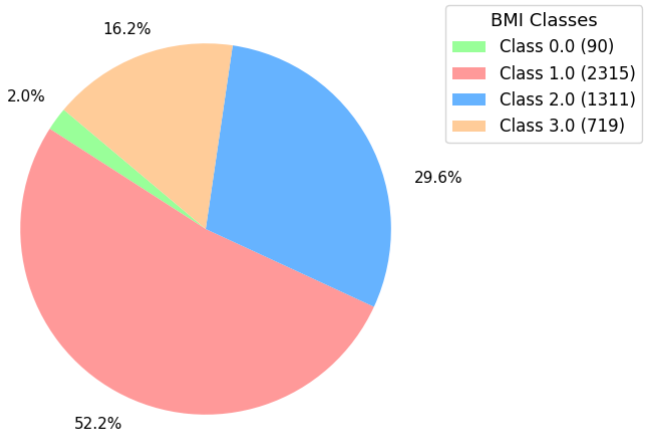


Figure 4.4 – Records per class after the missing data cleanse.

After this preprocessing thecnique, the features values distribution per BMI class still follows standard BMI categorization, seen in Figure 4.5. Where the overall average value of the features gradually increases as the BMI classification is higher. The outliers weren't treated because the dataset undergoes class merging in the next step, combining class 0 with class 1, and class 2 with class 3. This work is planned to be applied on binary classification problems, and therefore the target in this dataset is the BMIclass since it represents categorical data. Still, the BMIclass represents a multi-class problem, and for that reason the 4 classes were merged in 2 classes only. This class aggregation process is done by grouping the class 0 with 1, and the class 2 with 3, resulting respectively in classes 0 and 1. Therefore the new dataset problem corresponds to the identification of individual bodyweight higher than the normal weight.

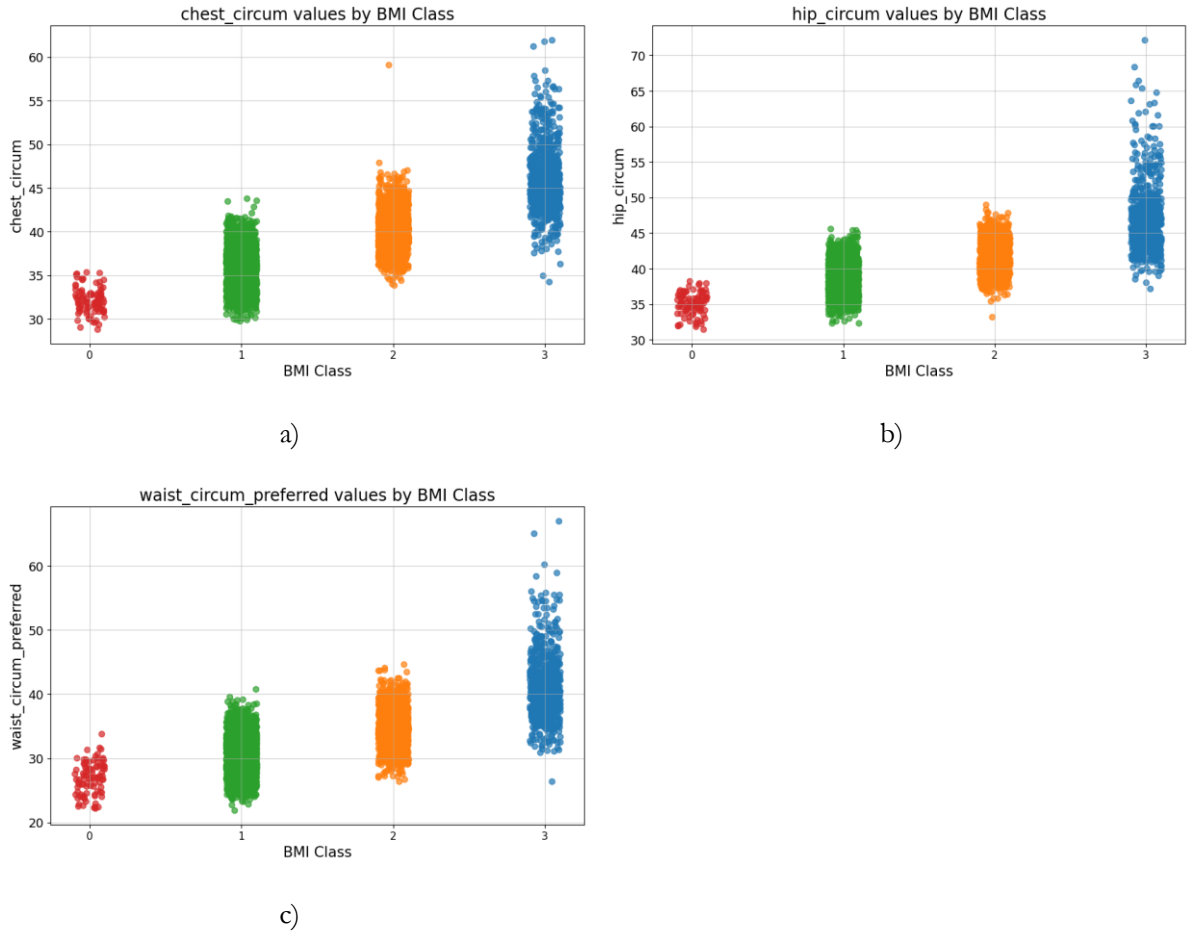


Figure 4.5 – Feature values distribution per BMI class: a) chest circumference; b) hip circumference; c) waist circumference.

The BMI feature was removed since the body mass index shall now be classificated based on body characteristics, and not calculated a specific BMI value. Now with a binary problem after the class merging, there is still a significant class overlap, observable in Figure 4.6, which usually takes place on real-world scenarios and provides different types of data points to be tested by the methods to be developed.

In order to understand the relationships between the different variables present in the data, the correlation matrix was used. This matrix displays the correlation coefficient for all the possible pairs of variables, ranging between [-1,1]. A positive relationship is stronger the closer the value is to 1, meaning that as one variable increases the other tends to increase as well. On the contrary, when a variable tends to increase as the other decreases, this negative relationship is stronger the closer it is to -1. The closer the value is to 0, the weaker the relationship between the variables.

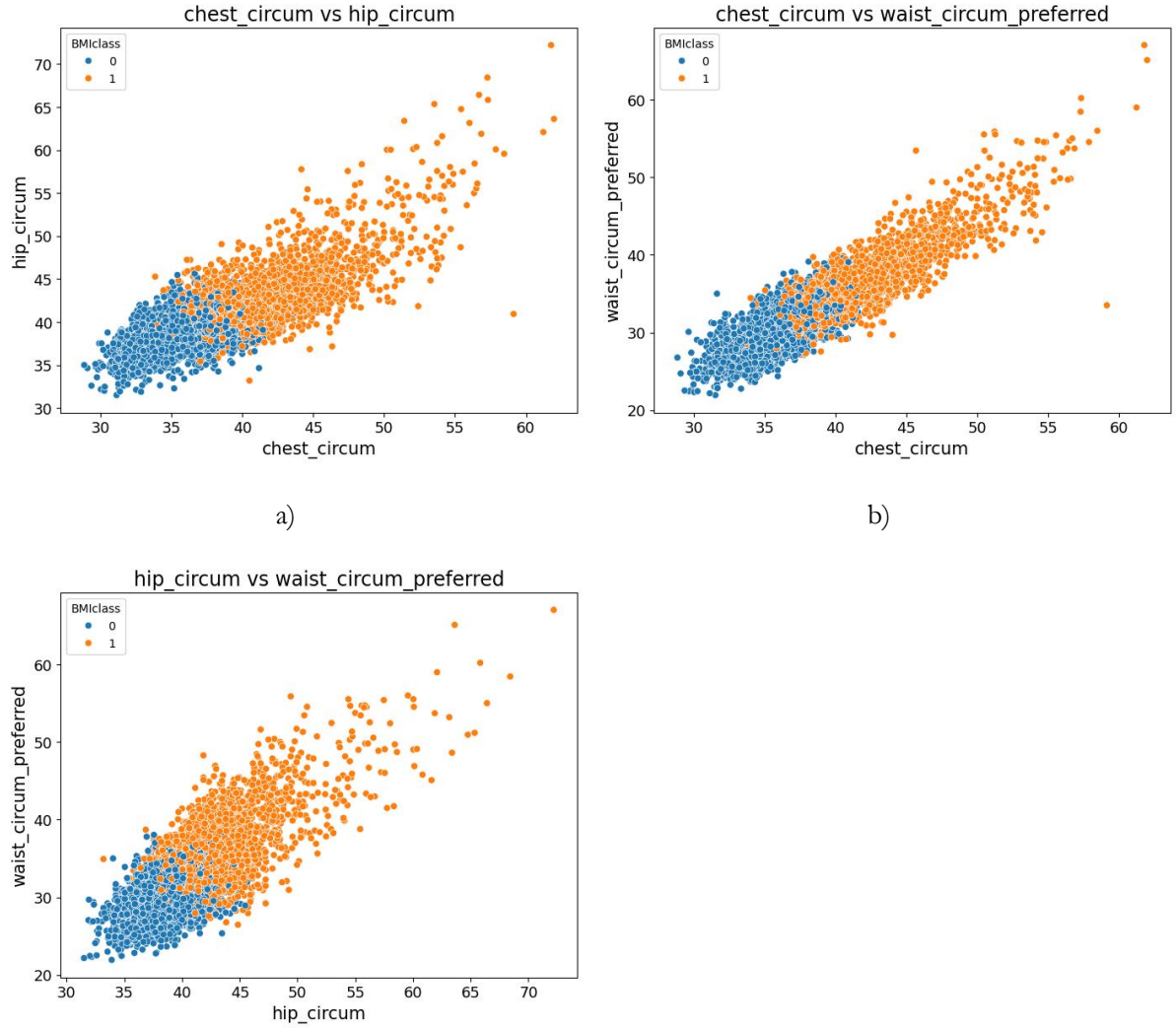


Figure 4.6 – Data points distribution per class across all pairwise feature combination: a) chest circumference and hip circumference; b) chest circumference and waist circumference; c) hip circumference and waist circumference.

The correlation matrix is displayed in Figure 4.7. Observing the correlation matrix, we can conclude that all features have a significant positive association with BMI classification, showing strong correlations between body measurements. As expected, this positive correlation signifies that higher BMI values tend to have larger body measurements.

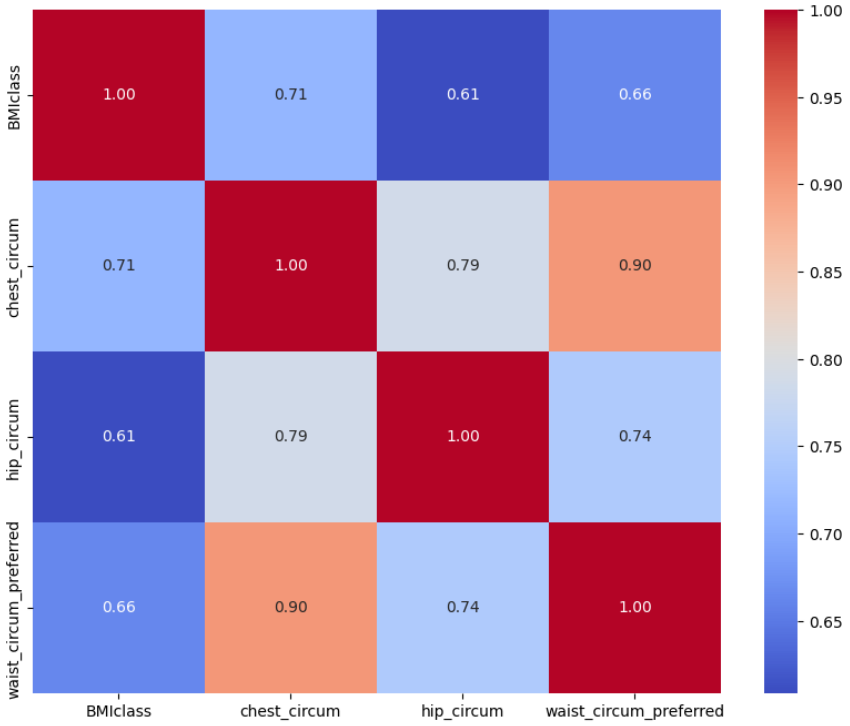


Figure 4.7 – BMI dataset correlation matrix.

Figure 4.8 indicates the resulting dataset which will be used on the testing of the methods. This dataset now contains 4435 records, with 3 features (chest, waist and hip circumference) and 1 target: (BMI class).

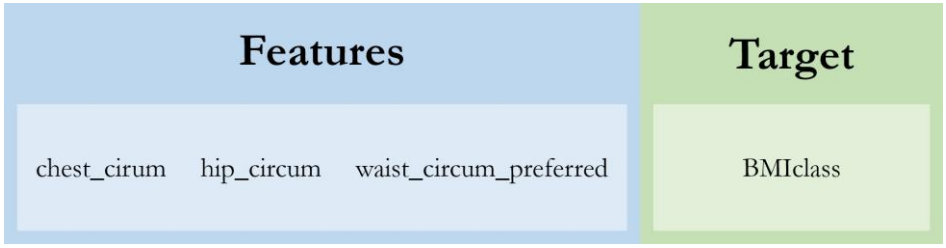


Figure 4.8 – Format of the final BMI dataset.

4.2.2 Cardiovascular Risk Assessment

The Cardiovascular Risk Assessment (CRA) dataset is a real-world dataset, provided by CHUC – Coimbra Hospital and University Centre. This medical institution in Portugal is recognized by its highly qualified reference centers in medical and surgical specialized areas. Its main mission is to deliver high-quality clinical care in a context of pre and postgraduate training with a strong focus on research, scientific knowledge and innovation.

The CRA dataset was delivered in an excel sheet containing data on 1544 patients, admitted to the Cardiology ICU at CHUC with myocardial infraction, between 2009 and 2016. The study received prior approval, and all patient information was

anonymized to ensure privacy. For these patients, the doctors used the GRACE tool to calculate their 6-month mortality risk, in the context of secondary prevention aimed at dealing with patients already diagnosed with acute coronary syndrome. All data modifications were done following previous medical indications.

The complete data consists of 145 columns, but only 8 features are essential to the 6-month mortality prediction, as they align with the risk factors used in the GRACE score. From these, only 6 features were present in the dataset: age, heart rate, systolic blood pressure, serum creatinine, Killip class and STEMI. The remaining features (cardiac arrest and elevated cardiac enzyme/marker) needed to be calculated based on existential data.

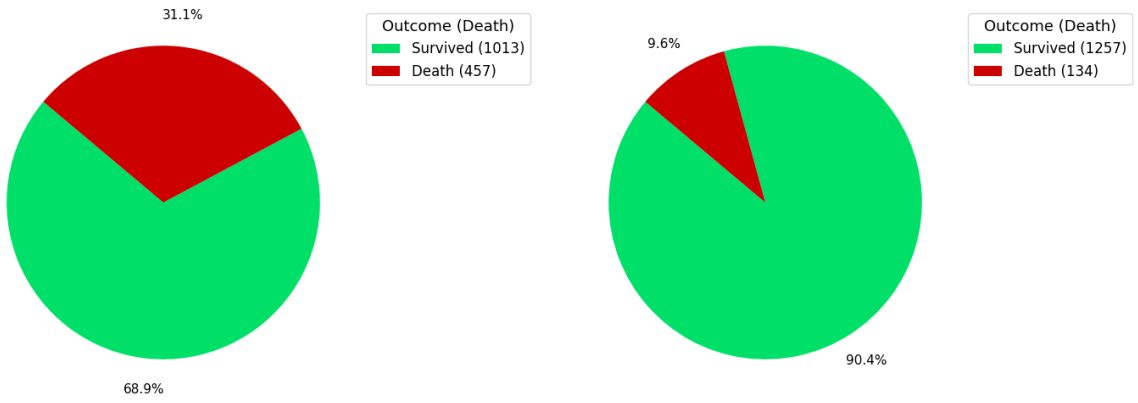
The cardiac arrest feature was calculated based on the Killip feature. If the Killip feature has the value 1, 2 or 3 then the cardiac arrest shall be 0. On the other hand, if the Killip feature has the value 4 then the cardiac arrest shall be 1. The elevated cardiac enzyme/marker used was troponin, which is a protein found in certain types of muscle in the body, such as the heart muscle, that is released into the bloodstream when the heart is injured or damaged. The troponin feature was calculated based on the type of acute coronary syndrome feature (ACS). If the type of ACS has the value 0, 1 or 2 then the troponin shall be 1. On the contrary, if the type of ACS has the value 3 then the troponin shall be 0. Table 4.2 indicates the selected 8 features and target to be used from the dataset.

Table 4.2 – CRA Dataset feature description.

Feature	Description	Data type	Unique values	Missing data
Age	Patient's age	Numeric	65	0
Heart rate	Patient's heart rate at admission	Numeric	105	22
Systolic blood pressure	Patient's systolic blood pressure at admission	Numeric	147	20
Serum creatinine	Patient's maximum creatinine level recorded	Numeric	929	21
Killip class	Patient's maximum killip class recorded	Numeric	4	3
STEMI	Indicator of STEMI	Binary	2	1
Cardiac arrest	Indicator of whether the patient had a cardiac arrest	Binary	2	0
Troponin	Indicator of troponin presence in patient's blood	Binary	2	0

Some specific modifications in the data needed to be done to align with medical considerations, due to typographical errors. These included removing a patient's record with both heart rate and systolic blood pressure at 0, the change of a patient's death date of the year 2007 to 2017, and a patient's heart rate of 11 to 71.

The binary target variable, death, contained 74 records of missing data. This variable indicates whether a patient survived or died within a six-month time window following hospital admission, with two possible outcomes: '0' for survival and '1' for death. All rows with missing data from features and target variables were removed, resulting in a total of 1430 records on patients. Among the patients who died (death = 1) some records were adjusted to ensure the target accurately reflects the 6-month mortality risk. In order to do so among the deceased patients, if the time between the admission date and the death date exceed 6 months, the patient's death record was changed to 0, meaning the patient did not have a 6-month mortality. Otherwise, if the difference was lower than 6 months, the patient's death record remained unaltered. For that matter, 38 records of deceased patients with no information on death date were deleted. Figure 4.9a and Figure 4.9b depicts the quantity of records per death case before and after the preprocessing, respectively.



a) b)

Figure 4.9 – Records per class a) before and b) after preprocessing.

With the aim of understanding how different features are interrelated and affect the death outcome, a correlation matrix was used. Remembering a value close to 1 indicates a strong positive relationship, a value tending to 0 indicates a little to no relationship, and a value closer to -1 indicates a strong negative relationship, the correlation matrix is displayed in Figure 4.10.

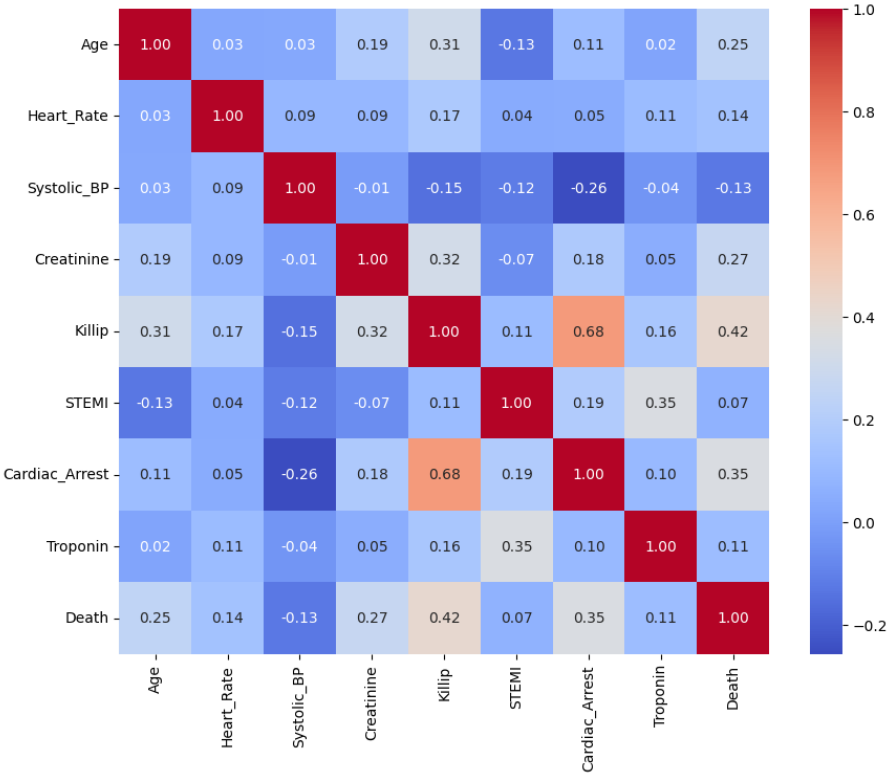


Figure 4.10 – CRA dataset correlation matrix.

With the correlation matrix we can observe that most features have a weak to moderate correlation between themselves, showing that they are relatively independent from each other. The death outcome seems to be the most correlated to killip (0.42) and cardiac arrest (0.35), followed by creatinine (0.27) and age (0.25). This indicates that patients with a higher killip class and who suffered a cardiac arrest have a higher risk of death. These two features are the most correlated with death outcomes, likely because both are related in general clinical context, verified by the strong positive correlation presented amongst them (0.68). The results also show that patients with higher creatinine levels and increased age are associated with a higher risk of death, which aligns with expectations. Figure 4.11 indicates the resulting dataset which will be applied on the developed methods. This dataset now contains 1391 records, across 8 features: age, heart rate, systolic blood pressure, serum creatinine, killip class, STEMI, cardiac arrest at admission, troponin presence and 1 target: death outcome.

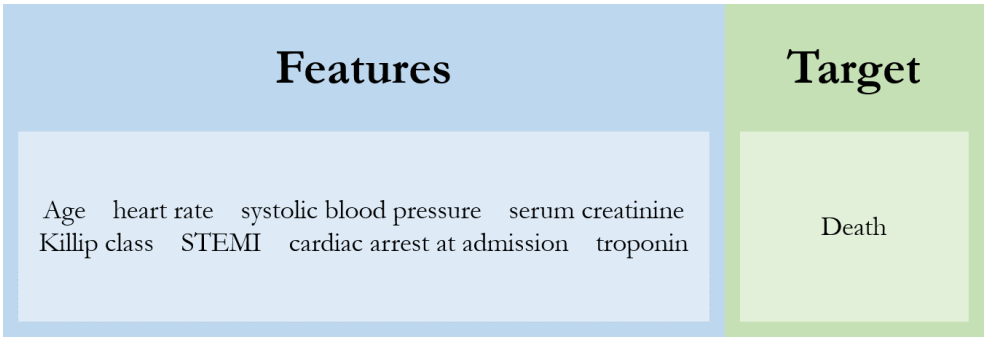


Figure 4.11 – Format of the final CRA dataset.

4.3 ML Model

Based on the previously mentioned datasets, a ML model was developed to estimate the risk of death within six months' post-admission. The ML model selection was not a fundamental focus, because the developed methods are meant to be model independent. Nevertheless, to reflect a real case scenario some choices were set.

Usually, ML models are refined to have the highest possible metric values, especially on medical risk scenarios where incorrect predictions can have the most severe aftereffects, such as deaths. However, so that the developed methods can effectively be evaluated with a decent amount of incorrect predictions while still reflecting the high evaluation metrics used on such environments, a goal of near 80% average on all evaluation metrics (accuracy, sensitivity, specificity and AUC) was set for the ML model choosing. Each dataset used a different model for the purpose of, not only achieve these evaluation metric goals based on the respective data, but also for the verification that the developed methods are model agnostic.

In spite of that, the two datasets shared some preprocessing steps. In both, 70% of the data was used for training and the remaining 30% for testing. Also, all features were normalized in the range [0, 1] with *MinMaxScaler* using the *scikit-learn* library, so that all features have the same ranges. Once the data was scaled, the SMOTE method was applied to the training data to address any class imbalance. SMOTE is an oversampling method, that generates randomly new instances of the minority class by combining each minority class sample with its nearest neighbors of the same class [59]. These instances are created based on the features of the original dataset so that they become similar to the original instances of the minority class [60].

4.3.1 Body Mass Index Dataset

For the BMI dataset, a Gaussian Naïve Bayes model was found to be good for the characteristics of the data. Gaussian Naïve Bayes is a probabilistic classifier, calculating the probability of a class given a set of features. It is a variant of the Naïve Bayes algorithm, employed for continuous data and assuming the features within each class follow a Gaussian distribution. The likelihoods of the features conditioned

on the classes are represented by a Gaussian probability density function, defined by equation (4.1), where X_i is the value of the feature *i*, C_j is the class labeled *j*, μ and σ^2 are respectively the mean and variance of each feature for each class calculated during training [61].

$$P(X_i \mid C_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{\frac{-(x_i - \mu_j)^2}{2\sigma_j^2}} \quad (4.1)$$

This model achieved an average score of 83.7% across all metrics, more specifically an accuracy of 84.1%, sensitivity of 70.6%, specificity of 96.5% and AUC of 83.6%. Sensitivity and specificity are metrics used to evaluate the true positive rate and the true negative rate, respectively. In a healthcare scenario, sensitivity represents the model's ability to correctly identify patients with a disease, and specificity represents the model's ability to correctly identify patients without the disease. With that said, a lower percentage of sensitivity is preferred compared with specificity, because it's the one which will present FN to be tested on the developed methods. And so, the difference observed between sensitivity and specificity does not seem to be troublesome for the methods testing.

The training instances consisted of 3104 data points, equally divided per class due to SMOTE application. The test instances consisted of 1331 data points, 695 points belonging to class 0 and the remaining 636 to class 1.

4.3.2 Cardiovascular Risk Assessment Dataset

For the CRA dataset, a Support Vector Machine (SVM) was established as good model for the specific data. SVM is a maximization/minimization algorithm used to identify the optimal linear separable hyperplane, defined by (4.2) where w is the coefficient vector and b is a constant, to separate data points of different classes. The hyperplane is computed until a margin is maximized, equivalent to (4.3), representing the distance between the positive class and the negative class [62].

$$f(x) = \langle w, x \rangle + b \quad (4.2)$$

$$\min \frac{1}{2} \|w\|^2 \quad (4.3)$$

This model achieved an average score of 79.3% across all metrics, more specifically an accuracy of 77.8%, sensitivity of 82.3%, specificity of 77.2% and AUC of 79.9%. The training points consisted of 1764 data points, equally divided per class due to

SMOTE application. The test points consisted of 418 data points, 372 points belonging to class 0 and the remaining 46 to class 1.

4.4 Methods

All methods come from the premise that, given a set of instances from a dataset $D = \{x_1,\ldots,x_N\}$ where N is the number of data instances in D, a model A outputs $A(x) \in \{0,1\}$. For each point in D, a known label $T(x) \in \{0,1\}$ is available. The certainty, $C(x)$, of the ML output, $A(x)$, is computed in the interval [0-1], where a value of 0.0 means complete lack of certainty on the output, and a value of 1.0 means total certainty on the output.

The following subsections describe the implemented methods to measure the trustworthiness of a ML prediction. The hyper parameter tuning will be commented on chapter 1.

4.4.1 Data proximity based

This method is based on the Intuitive Certainty Metric (ICM) proposed by [56]. It uses a distance function d to interpolate the performance of A for new data points, relying on radial basis functions, specifically a Gaussian function centered on the new data point. This way, the training points closer to the new data point have more influence in the prediction outcome than the training points further from it. The definition of a near training point is established through a parameter, standard deviation $\sigma \in [0; \infty]$, which manages the neighborhood size around the new data point. For a feature vector f, the ICM is calculated as shown in (4.4). ICM evaluates the certainty of f with a value in the range [0,1], based on how similar it is to training points. It does this by dividing the positive contributions of training points over the contributions of all training points

$$\mathrm{ICM}(f|\sigma,D) = \frac{1}{Z(f|\sigma,D)} P(f|\sigma,D) \tag{4.4}$$

Equation (4.5) defines the contribution of training points with the same label as f. The sum runs over all training instances whose match the label of f. The σ controls the width of the Gaussian exponential function. A smaller σ makes the function give higher contributions to nearby points (smaller distances), while a larger σ considers distant points as similar

$$P(f|\sigma,D) = \sum_{x \in D(T(x)=A(f))} \exp\left(-\frac{d(f|x)^2}{2\sigma^2}\right) \tag{4.5}$$

Equation (4.6) defines the normalization constraint. It operates similarly to equation (4.5), except that the sum runs over all training instances, regardless of their label.

$$Z(f|\sigma,D)=\sum_{x \in D} \exp\left(-\frac{d(f|x)^2}{2\sigma^2}\right) \quad (4.6)$$

This section of the ICM is the basis of the Data Proximity-based method developed in this work.

There are 6 arguments passed to the developed certainty function upon calling it: x^*, the instance of which the certainty is going to be calculated; $A(x^*)$, the predicted class label for the point; σ_0, the deviation value to be used in case $A(x^*) = 0$. σ_1, the deviation value to be used in case $A(x^*) = 1$; D, the data space to be compared with the point, which is the data used to train the model A; $T(D)$, as well the label of each point in D.

Figure 4.12 describes the pseudocode for Data Proximity-based method measuring the certainty of a point's prediction, considering a feature vector f. For each point in D, x, the distance between f and x is calculated based on a distance function d. If $A(f)$ is the same as $T(x)$ the contribution of x is calculated and added to P. Despite the relation of $A(f)$ with $T(x)$, the contribution of x is always calculated and added to Z. Finally, the certainty is calculated by dividing P with Z.

```
- DPBMeasure(f, A(f), σ0, σ1, D, T(D)):
      If A(f) is 1:
         Set σ = σ1
      Else:
         Set σ = σ0
      For j = 1..len(D)
         distance = Distance(f,Dj) % Measure distance between points
         If A(f) is y(Dj):
            Add exp(-distance² / 2σ²) to P
         Add exp(-distance² / 2σ²) to Z
      Add epsilon to Z % Add small value to avoid divisions by zero
      Certainty = P / Z
Return: Certainty
```

Figure 4.12 - Data Proximity-based method pseudocode.

The epsilon factor is a small positive constant, and was added in order to prevent undefined values upon calculating the Certainty. Undefined certainty values might happen due to divisions by 0, or exceptionally small values. These specific cases happen when all the distances between f and D_j are considerably large, resulting in small exponential values, and therefore a normalization constant Z, of or very close to 0. To achieve an epsilon value which prevents this problem while not having a strong impact on the final results of the method, different small values were tested

and analyzed the changes of their impact on the method's results. An epsilon value of 1 x 10^{-10} was found to be enough to prevent undefined certainty values, without influencing the final results.

The distance function *d* was adapted differently for each of the two datasets used in this study. This choice was made due to the data types that each dataset contains. For the BMI dataset, the Euclidean distance was chosen because its commonly used for low dimensionalities with same scale variances and numerical data. And for the CRA dataset, the Gower distance was chosen because it's the most suited for cases with mixed types of data, such as quantitative and categorical data.

4.4.2 Clustering based

This method is based on the assumption that the prediction for an instance x^* is trustworthy if x^* is close to a cluster of training points with the same class as the class predicted for x^* [43]. Clustering refers to the task of identifying groups or clusters in a data set [63]. And so, for the purpose of obtaining the odds of a test point belonging to each cluster, the soft clustering method Fuzzy C-Means (FCM) was put into practice.

The FCM is an iterative algorithm which provides a membership degree ∈ [0-1] for each cluster (*C*) to a given test instance, where the summation of each membership is equal to a total of 1. It starts by randomly initializing the clusters centers (v_j) by assigning random values within the data range. The membership values (u_{ij}) of data points to each cluster are then computed using (4.7), according to a fuzziness parameter (*m*) that determines the extent to which data points can be assigned to multiple clusters. The greater the value of *m*, the more mixed the memberships become.

$$u_{ij} = \left(\sum_{k=1}^{C} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{(m-1)}} \right)^{-1} \tag{4.7}$$

It is ensured that the sum of memberships for each data point across all clusters equals 1, as specified in (4.8), and $u_{ij} \in [0,1]$ to guarantee memberships in a percentage ratio.

$$\sum_{j=1}^{c} u_{ij} = 1 \text{ for all } i \tag{4.8}$$

Then, the cluster centers are recalculated based on the updated membership values, according to (4.9):

$$v_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{4.9}$$

These last two steps are repeated until convergence of either the objective function, as in (4.10), or until the changes in the clusters centers or the membership values are bellow a threshold (ε), as in (4.11), or a maximum number of iterations is reached.

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - v_j \right\|^2 \tag{4.10}$$

$$\left\| u^{p} - u^{p-1} \right\| < \varepsilon \tag{4.11}$$

The certainty calculation of a prediction is the membership value assigned to the cluster with the most training points of the same class as the predicted, since it is already on a range of [0-1]. To determine which cluster contains the majority of each class, it was identified the cluster with highest membership value for each training point. And then for each cluster, calculated how many training points of each class belong to it.

Figure 4.13 describes the pseudo-algorithm for Clustering-based method measuring the certainty of a point's prediction.

- **CBMeasure()**:
  Given a training set $S \in Rd$, where d is the number of features
  Given a point $x \in Rd$, and a model A outputs $A(x) \in \{0,1\}$
  1. r = *ClusterCreation*(*S*.T) % Create clusters
  2. membership = *ClusterPrediction*(*x*.T, r) % Return the probability to each cluster
  3. **If** $A(x)$ is 0:
  4. certainty = membership[0]
  5. **Else** $A(x)$ is 1:
  6. Certainty = membership[1]

Return: Certainty

Figure 4.13 – Clustering-based method pseudocode

The FCM clustering algorithm was implemented in Python using the *scikit-fuzzy* library. Since this method is going to be applied on binary classifications, a maximum number of clusters $C = 2$ was chosen and created with *skfuzzy.cluster.cmeans* function. This function used the transpose of the training data for the creation of the clusters, without any initial fuzzy c-partitioned matrix, so the algorithm is randomly initialized. Also, a fuzziness parameter $m = 2$ to control the degree of cluster overlap, a stop criteria threshold ε of 0.005 and 10000 as maximum number of iterations. The cluster membership prediction for a given test point, is accomplished using the *skfuzzy.cluster.cmeans.predict* function with the same parameter values of the clusters creation function.

4.4.3 Convex hull based

The convex hull of a set of points is the smallest convex set that contains the points [64]. Therefore, a convex hull formed by the training points of a specific class, is a convex set that encompasses all the training points of that class. Figure 4.14 illustrates a convex hull construction process in 2D, where the input data consists of a set of instances, and the result is the smallest possible polygon that fully encloses all the input data.

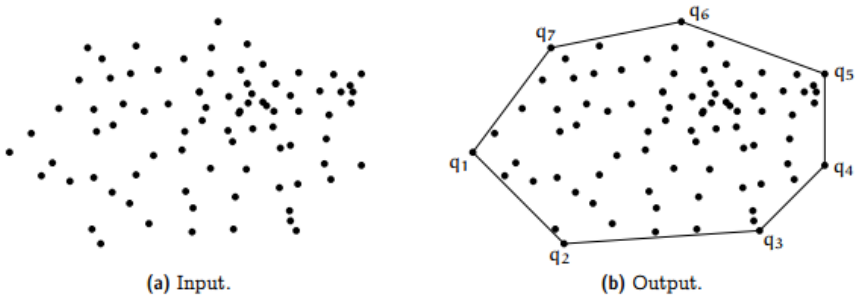


Figure 4.14 - Example of a convex hull in 2D space

The computation of the convex hull was made with resource to the *ConvexHull* class from the *SciPy* library, implemented with *Qhull*. Initially, the creation of one convex hull molded by all the training points was done. two convex hulls were created, one for each class of training points. In the following attempt, only one convex hull was originated, from the training points of the positive class (class 1), which yielded the best results, seen in chapters 5.1.3 and 5.2.3. This work uses 3D and 8D datasets (BMI dataset and CRA dataset, respectively), but computing convex hulls in high dimensions is a computationally difficult task [65]. Nevertheless, the distance of a point to a high (*col*) dimensional convex hull, can be approximated with the distance to hulls of multiple projections to lower dimensional spaces [43] Thus, multiple 3D projections of the original *col* - dimensional convex hull, formed by training points of class 1, were generated by combining sets of three features at a time. The projections were created by selecting specific features (1st, 2nd and 3rd features,

followed by 2nd, 3rd and 4th, and so on, until the set of *col* th, 1st and 2nd features were considered).

Because convex hulls are only useful to analyze the continuous distribution through a feature, the binary and categorical features were excluded of the computation of any convex hull due to their absence of depth. Figure 4.15 describes the pseudocode for the convex hulls creation procedure, which returns a list containing all the 3D convex hulls projections.

- **CreateCH()**:
  Given a training set $S \in Rd$, where d is the number of features
  Given a number $col \in d$
    1. **For** $j = 0..col$
    2. $S_j = S[j, j+1, j+2]$ % Select three consecutive columns
    3. hulls = $ConvexHull(S_j)$ % Protject the hull in three-dimension space
    4. **Append** hulls **to** hull

Return: hull

Figure 4.15 - Projected Convex Hull creation pseudocode

This method is based on the usage of the convex hull's geometric characteristics formed by training points, to determine whether the classification of a test point is the outcome of interpolation or extrapolation [43]. Interpolation occurs when a point lies within the convex hull's volume, as opposed to an extrapolation. With the center of the convex hull representing the place where the point is the most interpolated, transitioning to extrapolation at the convex hull's borders. For each 3D projected hull, the distance (d) between it and the project point (x_j) is calculated, and at the end the maximum distance (d_{max}) is reported as the final approximation distance of the point to the original convex hull. The distance to a convex hull was calculated with the support of the *hmesh* module from the *pygel3d* library. The provided distance value can take 3 different meanings:

- $d < 0$, for when the point is inside the convex hull;
- $d = 0$, for when the point lies exactly on the surface of the convex hull;
- $d > 0$, for when the point is outside the convex hull.

After the maximum distance d_{max} over all projected hulls is obtained, the certainty measurement of a prediction is calculated accordingly with the predicted class for the point. If the predicted class is the same as the convex hull's training points class (class 1), then the certainty is determined using the formulas (4.12). On the other hand, if the predicted class differs, the certainty is determined using the formulas (4.13).

$$\text{certainty}_{CH} = \begin{cases} \dfrac{|d_{max}|}{\alpha + |d_{max}|}, & \textit{if } x \textit{ is inside the hull (a)} \\ \dfrac{\beta}{\beta + d_{max}}, & \textit{if } x \textit{ is outside the hull (b)} \end{cases} \quad (4.12)$$

$$\text{certainty}_{CH} = \begin{cases} \dfrac{\gamma}{\gamma + |d_{max}|}, & \textit{if } x \textit{ is inside the hull (a)} \\ \dfrac{d_{max}}{\beta + d_{max}}, & \textit{if } x \textit{ is outside the hull (b)} \end{cases} \quad (4.13)$$

Figure 4.16 describes the pseudocode for the Convex Hull-based method measuring the certainty of a point's prediction. Due to issues of numerical precision upon calculating *d*, a small positive value (*tolerance*) of 1 x 10^{-12} was used instead of the value 0. Such issues are deviations due to rounding errors, which would cause some convex hull's vertex to possibly be identified as not belonging to the convex hull. This ensures that if a value *d* is very close to 0, but not exactly 0, it will still be considered part of the hull.

- **CHMeasure()**:
  Given a point $x \in Rd$, and a model A outputs $A(x) \in \{0,1\}$
  Given a small, positive value *tolerance*
  1. $d_{max} = -\infty$
  2. **For** $j = 0..col$
  3. $x_j = x[j, j+1, j+2]$ % Select three consecutive columns
  4. $d = dist(\text{hull}_j, x_j)$ % Calculate distance to convex hull with *PyGel*
  5. **If** $d > d_{max}$:
  6. $d_{max} = d$ % Return the maximum distance to hull
  7. **If** $A(x)$ is 1:
  8. **If** $d_{max} >$ *tolerance*: % Check if point is outside the hull
  9. certainty = (4.12)
  10. **Else**:
  11. certainty = (4.12)
  12. **If** $A(x)$ is 0:
  13. **If** $d_{max} >$ *tolerance*:
  14. certainty = (4.13)
  15. **Else**:
  16. certainty = (4.13)

Return: certainty

Figure 4.16 - Convex Hull-based method pseudocode.

4.4.4 Summary

The Data Proximity-based method analyses the distance between the test instance and all training instances with resource to Gaussian exponential functions, and user defined neighborhood size. The more similar the instances around the test instance are, the more trustworthy it shall be.

The Clustering-based method employs the Fuzzy C-means clustering algorithm, to assess the correlation of the test instance with each generated cluster. Ideally, this method assumes perfectly class separated clusters, allowing the measurement of the test instance's membership strength in relation to each class.

Finally, the Convex Hull-based method partitions high dimensional data into multiple 3D projections. It then calculates an approximated distance between the test instance and the convex hull. Similar to the clustering method, the ideal scenario is that each convex hull created corresponds to the polygon enclosing only training instances of a single class.

5 RESULTS

For observational purposes, the methods mentioned above were tested with the data instances from the test set. This approach was chosen to test the methods with a large number of data points, simulating a more realistic data scenario. This ensures that the results remain unaffected, as these data points are unknown to the ML model.

The results are presented with the distribution of correct and incorrect ML model predictions through the reliability levels, ranging from 0 to 1 as previously stated. Also, the Point Biserial Correlation is calculated in order to quantify the results. This coefficient was chosen due to its usage with the data types, since it needs continuous and binary variables [66]. It is a correlation coefficient between the correct predictions (metric variable) and its corresponding reliabilities (dichotomous variable), varying between [-1, 0] negative correlation and [0,1] positive correlation.

The following sections present the results obtained by the previously mentioned methods for each dataset, as well the interpretation of the same.

5.1 Body Mass Index Dataset

This section presents the results obtained using the Body Mass Index dataset on each of the proposed methods.

5.1.1 Data Proximity-based method

The application of the Data Proximity-based method to the BMI dataset involves the tuning of some parameter values. To determine the size of the significant neighborhood around a data point, defined by the parameter deviation σ, the distances between the points of the same class were calculated, and the average of the 20% shortest distances was used as the deviation value for each class. Figure 5.1 and Figure 5.3 represent the results when using this deviation procedure, generating σ_0: 0.056 and σ_1: 0.071.

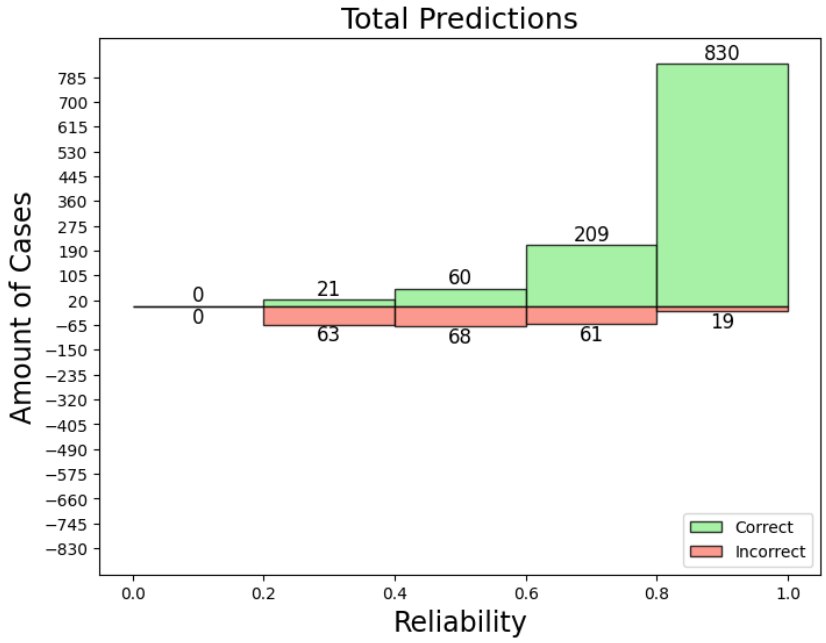


Figure 5.1 – Number of correct/incorrect predictions for each reliability interval using the DPB method, on BMI dataset; σ_0: 0.056; σ_1: 0.071

Figure 5.1 represents the amount of correct and incorrect predictions made by the model *A*. It can be seen that the majority of correctly predicted cases are located at the highest reliabilities (0.8-1.0), while incorrect predicted cases are less frequent at these levels compared to lower reliability levels. The fact that no predictions were assigned with reliability between 0.0 and 0.2, is result of two circumstances. Primarily, the lack of strong separability between classes shown in Figure 5.2 because points from the same class and points from the opposite class do not have very discrepant distances. For very large distances, the exponential function converges near 0, and for small distances, it converges near 1. Since the distances are never excessively large, the exponential decay function never produce values close to 0. Secondly, the high standard deviation for opposite class distances (std = 0.14) when compared to same class distances (std = 0.09) suggests a scatter of the points, meaning that some opposite class points are dispersed, leading to the lack of contribution for the reliability calculations. This can be worked around by tuning σ, in order to expand or enclose the size of the significant neighborhood around a data point.

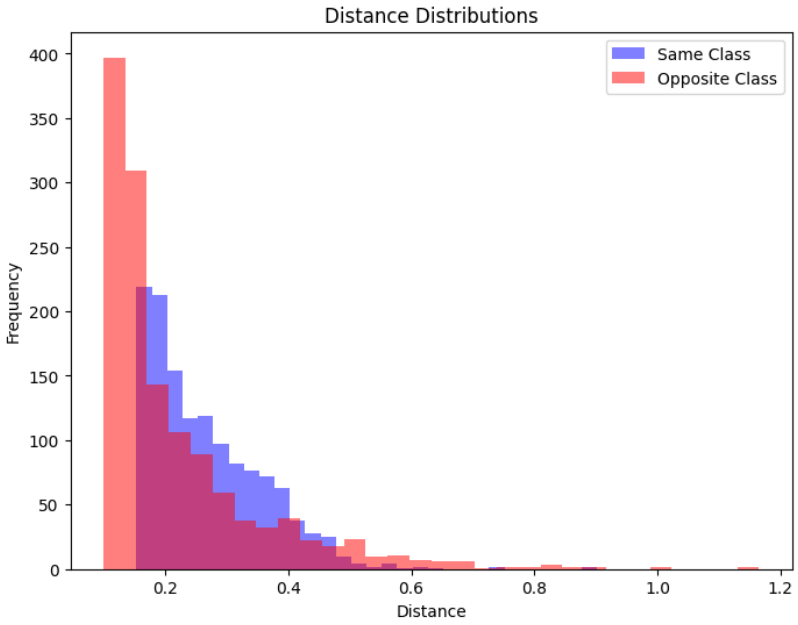


Figure 5.2 – Test points distance distributions to training points.

Figure 5.3 displays the percentage of each prediction type on reliability intervals, presenting more clearly the evolution of correct and incorrect predictions along the reliability scale. The results show an indubitably predominance of correctly predicted cases over the higher reliabilities (>0.5), with incorrectly predicted cases growing when moving towards the lowest reliabilities. The category in dominance changes exactly at the medium reliabilities [0.4, 0.6], where each type of prediction has almost a 50% chance of happening. It's the desired outcome because a reliability close to 50% should imply ambiguity, meaning the classification could result in any output with almost equal probability.

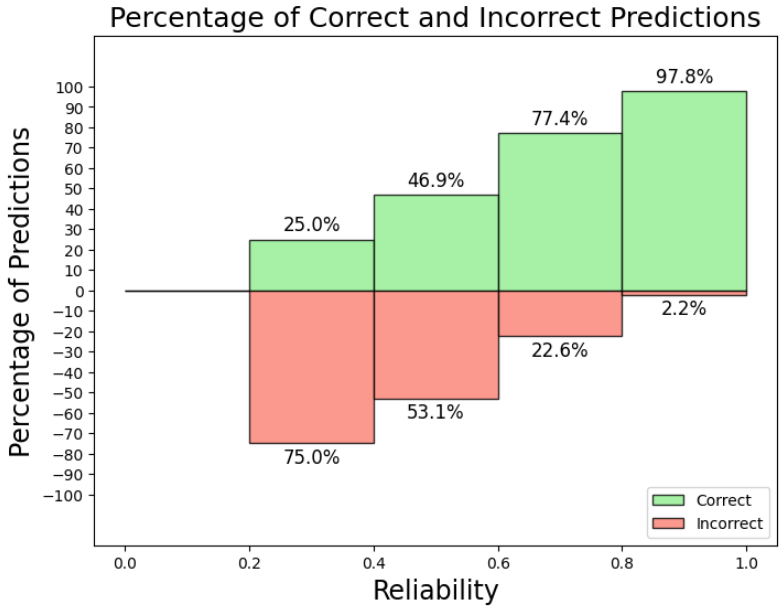
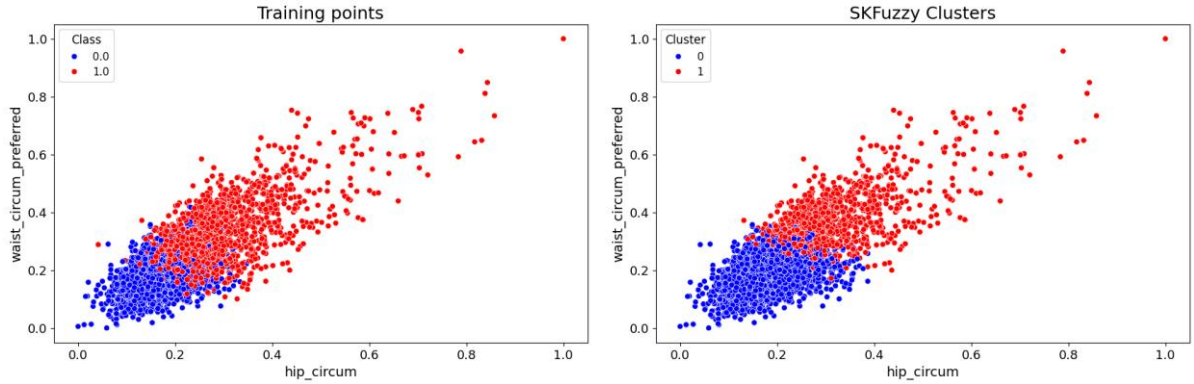


Figure 5.3 – Percentage number of correct/incorrect predictions for each reliability interval using the DPB method, on BMI dataset; σ_0: 0.056; σ_1: 0.071.

Applying the Point Biserial Correlation to the Data Proximity-based method using the BMI dataset, a coefficient value of 0.61 was obtained with a *p-value* = 8.55×10^{-136}. A correlation of 0.61 in a range of [-1,1] suggests a moderate to strong relationship between reliability values and prediction outcomes, and the extremely small *p-value* indicates that the correlation is statistically significant and not likely to have randomly happened.

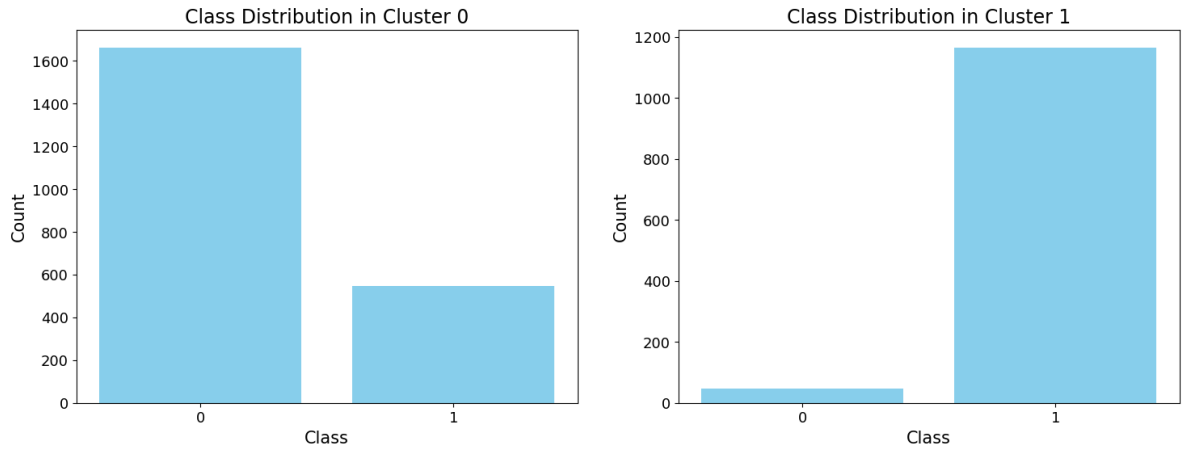
5.1.2 Clustering-based method

The application of the Clustering-based method to the BMI dataset was done with resource to the Fuzzy C-Means clustering algorithm, which used the following parameterization values, adjusted through experiments: C = 2, representing the number of clusters; m = 2, fuzzy parameter controlling how soft the clustering assignment is; *error* = 0.005, cluster centers position threshold; *maxiter* = 10000, maximum number of iterations; *init* = 'None', random initialization of cluster centers. The Figure 5.4a displays the training points distribution by class, and Figure 5.4b the data distribution of the points by each cluster per two features: *hip_circum* and *waist_circum_preferred*. The Figure 5.5 highlights the distribution of the classes across the two clusters formed using the training points.



a) b)

Figure 5.4 – BMI training points distribution: a) by class; b) by each cluster.



a) b)

Figure 5.5 – Number of BMI training points by class across the clusters: a) cluster 0; b) cluster 1.

Both figures demonstrate a moderate level of class blending in the clusters, especially in cluster 0. About 24% (545 data points) of the instances belonging to cluster 0 are from class 1, and only near 4% (48 data points) instances associated to cluster 1 are from class 0. This level of overlap represents a measurable degree of class mixing, which may impact the reliability calculation, since this method requires distinct class clusters. Figure 5.6 and Figure 5.7 present the results of the reliability measurement with the usage of these clusters.

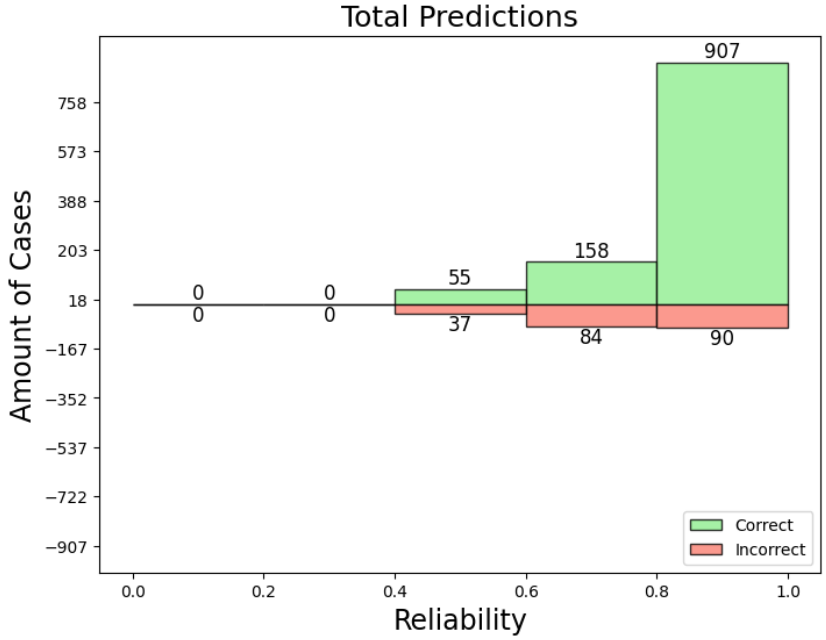


Figure 5.6 – Number of correct/incorrect predictions for each reliability interval using the CB method, on BMI dataset.

The amount of cases tends to rise with the reliability values, independently if it's wrong or right predictions. No cases are observed at low reliability values; instead, they start appearing only from mid-range reliabilities and increase until reaching a peak at the highest reliability levels. This suggests that the proximity between clusters may not be clearly defined. As a result, incorrect predictions may still be assigned moderate to high reliability scores, since they lie very close to their actual cluster. The two clusters centers distance between themselves about 0.3 in a Euclidean metric, which is a relatively small distance separation. This indicates the clusters are close together, causing them to overlap and having a weak separation. Besides this, the small class blending present in both clusters also contributes, at some degree, to the wrongful assignment of high reliability scores to incorrect predictions.

Figure 5.7 authenticates the previous phenomenon, where reliability scores start only at [0.4-0.6]. And even though the quantity of incorrect predictions is smaller within the overall of cases as the reliability increases, it was previously seen that in the enclosure of incorrect predictions only, its cases quantity grows. Proving to not be a very well-suited method for the purposes intended.

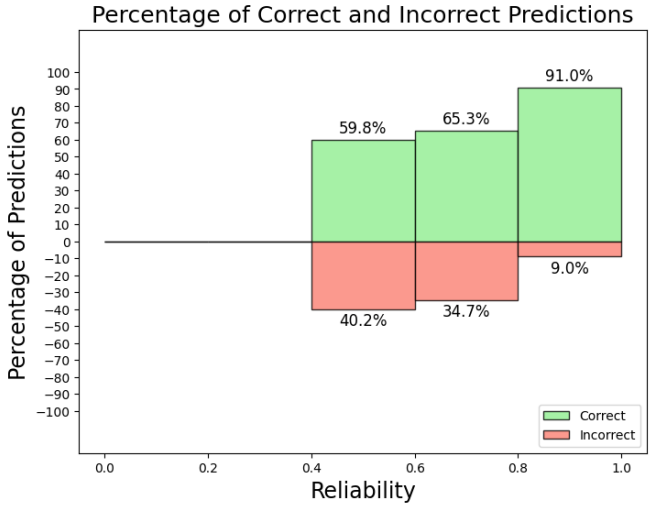


Figure 5.7 – Percentage number of correct/incorrect predictions for each reliability interval using the CB method, on BMI dataset.

These poor results are validated by the Point Biserial Correlation which, when applied to the Clustering-based method using the BMI dataset, a coefficient value of 0.34 is obtained with a *p-value* = 2.25×10^{-36}. A correlation of 0.34 in a range of [-1,1] suggests a weak relationship between reliability values and prediction outcomes, and the *p-value* indicates that the correlation is statistically significant, suggesting that the weak correlation observed is unlikely to have occurred by random chance alone.

5.1.3 Convex Hull-based method

The application of the Convex Hull-based method to the BMI dataset involves the tuning of parameters such as α, γ and β. Initially, the parameters α and γ were based on the average distance of the training points to the border of the convex hull, but

were adjusted through successive experiments, by analyzing the final results of each experiment. The parameter β is a value close to zero, so that there isn't much tolerance for when the points are outside the convex hull. Figure 5.8 and Figure 5.11 represent the results when using these parameters tunings, generating α: 0.02, γ: 0.01 and β: 0.002.

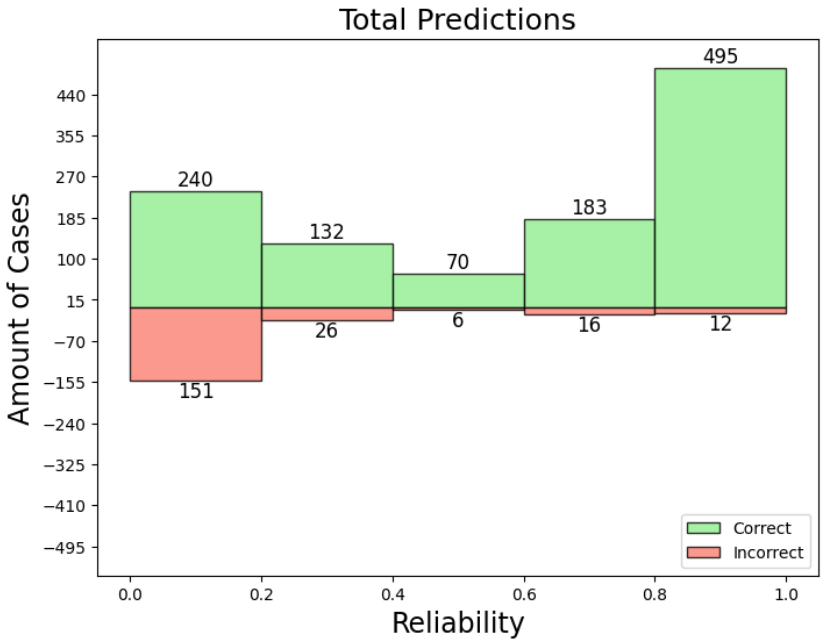


Figure 5.8 – Number of correct/incorrect predictions for each reliability interval using the CHB method, on BMI dataset; α: 0.02; γ:0.01; β:0.002.

Figure 5.8 represents the amount of correct and incorrect predictions corresponding to the predictions of the ML model *A*. The results don't show a clear increase of correct predictions with the reliability escalation. Transmitting a multimodal distribution idea, where there are almost two cluster cases around two different reliability scores, [0.0-0.2] and [0.8-1.0], with a lower value of 70 cases between them at [0.4-0.6] reliabilities. However, it's an indubitable fact that the majority of correctly predicted cases (495) are located at the highest reliabilities (0.8-1.0), while the majority of incorrectly predicted cases (151) are located at the lowest reliabilities (0.0-0.2).

The multimodal distribution effect happens due to class overlap, specifically the quantity of class 0 points contained within the convex hull formed by class 1 points. This is observable in Figure 5.9 using the test points of both classes, where many test points of class 0 have a distance to the convex hull ≤ 0.0, implying they are inside the convex hull. Due to the low value of α and γ, the class 0 points correctly classified by the model with $|d_{CH}| \geq \gamma$ will have a very low reliability value, and if incorrectly classified with $|d_{CH}| \geq \alpha$ will have higher reliability values. This is further noticed in Figure 5.10, where the majority of correctly classified examples with reliability ≤ 0.4 are class 0 test points.

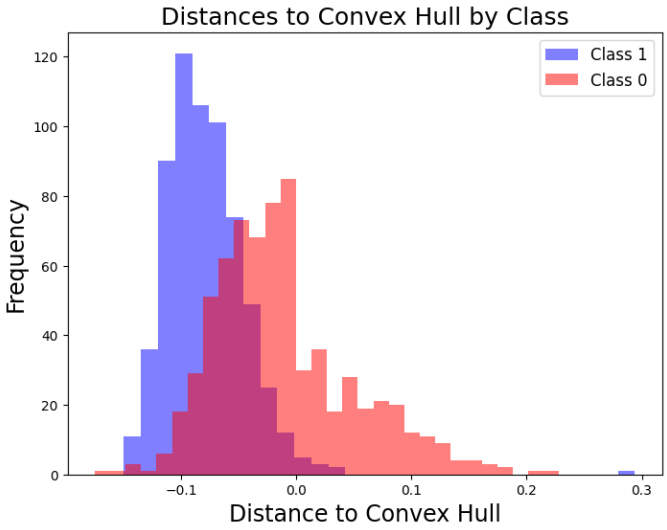


Figure 5.9 – Distance to convex hull by data points class.

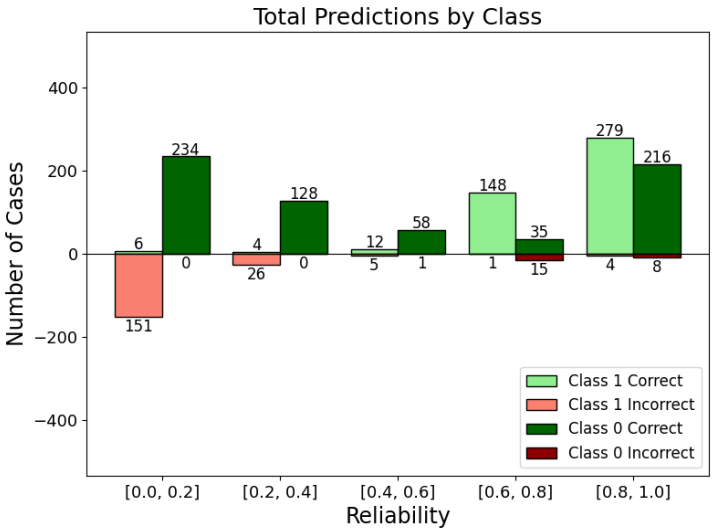


Figure 5.10 – Number of correct/incorrect predictions for each reliability interval by class.

Figure 5.11 shows the percentage of each prediction type on reliability intervals, presenting clearly the evolution of correct and incorrect predictions along the reliability scale. It's possible to perceive that the proportion of correctly classified predictions tends to impose to the proportion of incorrectly classified predictions as the reliability increases.

Applying the Point Biserial Correlation to the Convex Hull-based method using the BMI dataset, a coefficient value of 0.40 was obtained with a *p-value* = 9.27×10^{-52}. A correlation of 0.40 in a range of [-1,1] indicates a moderate relationship between reliability values and prediction outcomes. Additionally, the extremely small *p-value* indicates that the correlation is statistically significant, meaning it is unlikely to have occurred by chance.

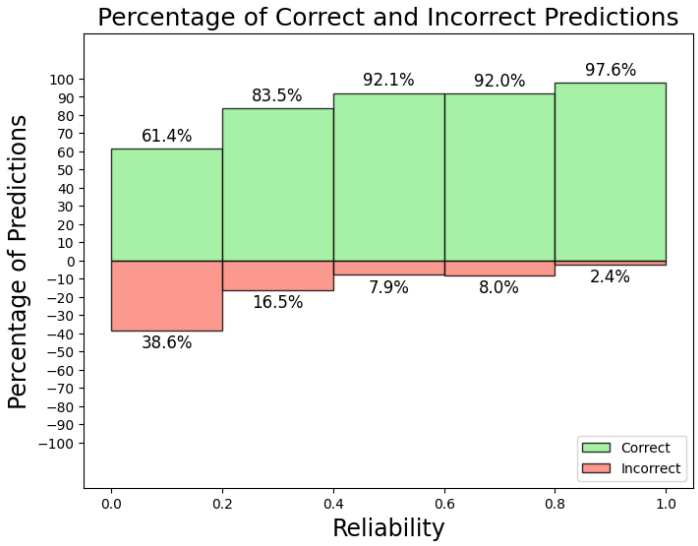
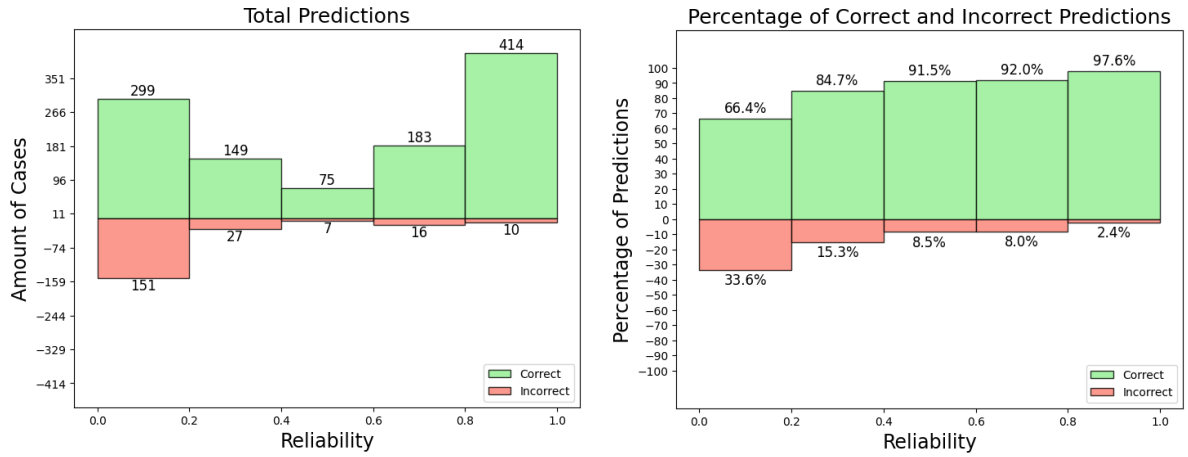


Figure 5.11 – Percentage number of correct/incorrect predictions for each reliability interval using the CHB method, on BMI dataset; α: 0.02; γ:0.01; β:0.002.

In order to verify the difference of a projected distance to the real distance, and since the BMI dataset is a 3D dataset, the same method was applied with a direct creation of the convex hull, using the 3 features. Figure 5.12 present the results for this experiment.



a) b)

Figure 5.12 – Results without projected hulls: a) Number of correct/incorrect; b) Percentage number of correct/incorrect.

The results for both convex hull creation procedures are very similar, proving that using projected hulls to obtain a projected distance is indeed a close approximation to the real distance, and so usable for convex hulls of higher dimensions, as is the case of the following dataset application.

5.2 Cardiovascular Risk Dataset

This section presents the results obtained using the Cardiovascular Risk dataset on each of the proposed methods.

5.2.1 Data Proximity-based method

The application of the Data Proximity-based method to the CRA dataset was done with the same procedures as on the application for the BMI dataset, previously stated. Figure 5.13 and Figure 5.14 represent the results when using these same procedures, generating σ_0: 0.056 and σ_1: 0.074.

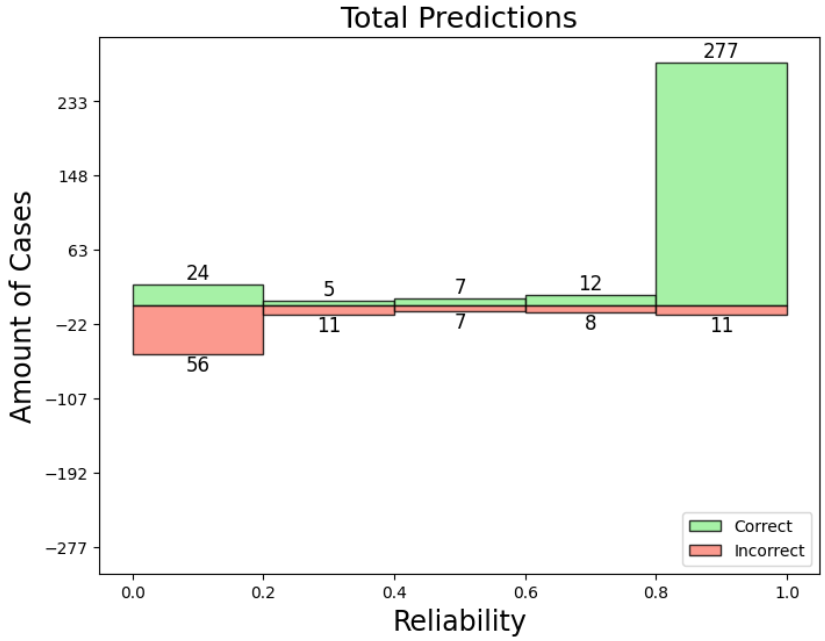


Figure 5.13 – Number of correct/incorrect predictions for each reliability interval using the DPB method, on CRA dataset; σ0: 0.056; σ1: 0.074.

Figure 5.13 shows that the majority of incorrectly classified cases are assigned to lower reliabilities between 0.0 and 0.2, while the majority of correctly classified cases are located at higher reliabilities, between 0.8 and 1.0 reliability. This is the desired outcome, because high reliabilities should correspond to correct predictions, while low reliability should indicate a higher likelihood of incorrect predictions.

The relatively small cluster of correct predictions between 0.0 and 0.2 reliabilities, is likely due to the quality of the dataset used. About 75% of the total points correctly predicted with reliability lower than 0.2, are from the minority class. These are points that are located closer to training points of the opposite class, than others with higher reliability. Since this is a 6-dimensional dataset, observing the distribution of points on a spatial space with the decision boundary is a difficult task.

However, the analysis of the positive contributions, *P*, and the negative contributions, *N*, gives an insight to the spatial distribution of training points near

the test point. *P* provides an idea of how influential (relative to distance and quantity of points) the class points of the same predicted class are, while *N* offers the same idea for the opposite class points.

Therefore, the values of *P* and *N* were determined for the points present in the *reliability* <= *0.2* ∧ *correct_prediction* = *1* set (set A), and for the points present in the *reliability* >= *0.8* ∧ *correct_prediction* = *1* set (set B). The average values of the set A were $P = 1.7991 \times 10^{-6}$ and $N = 0.0010$, the average values of the set B were $P = 0.0004$ and $N = 9.5929 \times 10^{-7}$. In set A, *P* is much smaller than *N*, indicating that the training points closest to the given test points are of the opposite class, resulting in a lack of reliability. In set B, *P* is significantly higher than *N*, hinting that the majority of points near the given test points are of the same class, causing higher reliability in the predictions.

Figure 5.14 reflects the already seen behavior in Figure 5.13: the higher levels of reliability are mostly restrained to correct predictions, gradually decreasing the quantity of incorrect predictions as the reliability is higher.

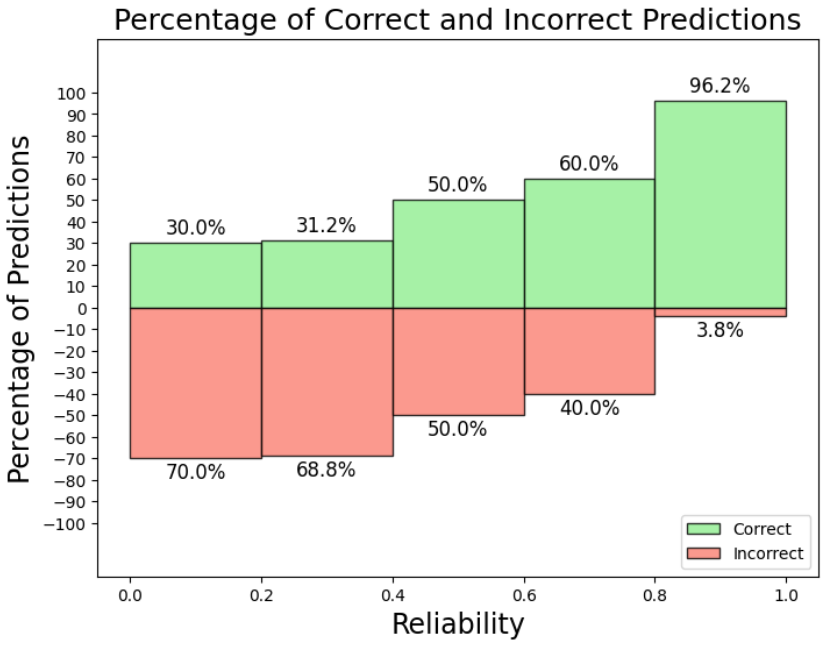


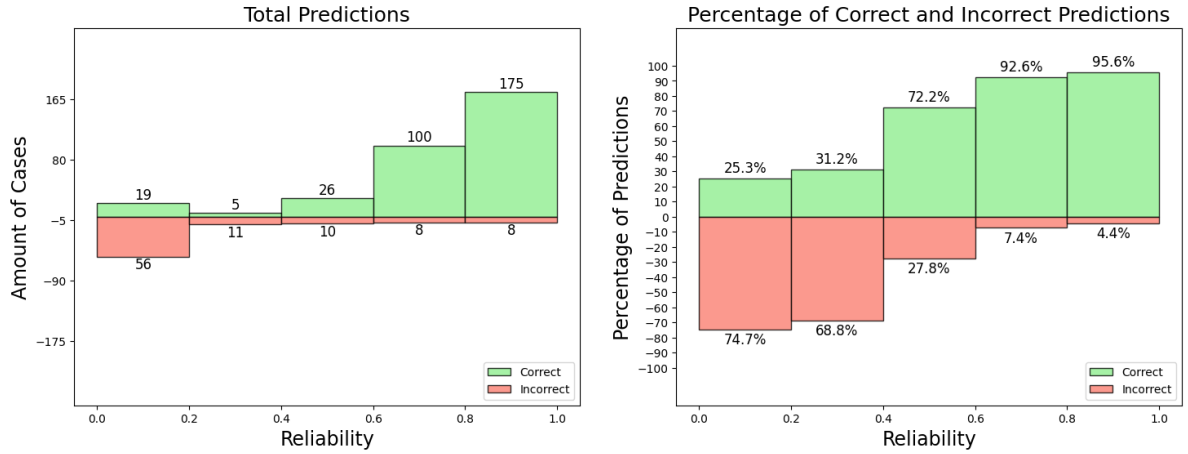
Figure 5.14 – Percentage number of correct/incorrect predictions for each reliability interval using the DPB method, on CRA dataset; σ0: 0.056; σ1: 0.074.

For reliability measures around 0.5, the percentages of correct and incorrect predictions are equal (50%). This is exactly where the majority of the prediction type changes, and it's an ideal scenario, because points without enough information to make a reliable prediction should lead to a near-random outcome.

Applying the Point Biserial Correlation to the Data Proximity-based method using the CRA dataset, a coefficient value of 0.66 was obtained with a *p-value* ≈ 0 (3.97×10^{-54}). A correlation of 0.66 in a range of [-1,1] suggests a meaningful relationship

between reliability values and prediction outcomes, and the extremely small *p-value* indicates that the correlation is statistically significant.

The following Figure 5.15 demonstrates the result of further experiments that were tested, in order to minimize the concentration on extreme levels of reliability and reduce quantity of incorrect predictions with high levels of reliability, by adjusting the parameters σ_0 to 0.139 and σ_1 to 0.074, with a coefficient value of 0.65 with a *p-value* ≈ 0 (6.71×10^{-51}).

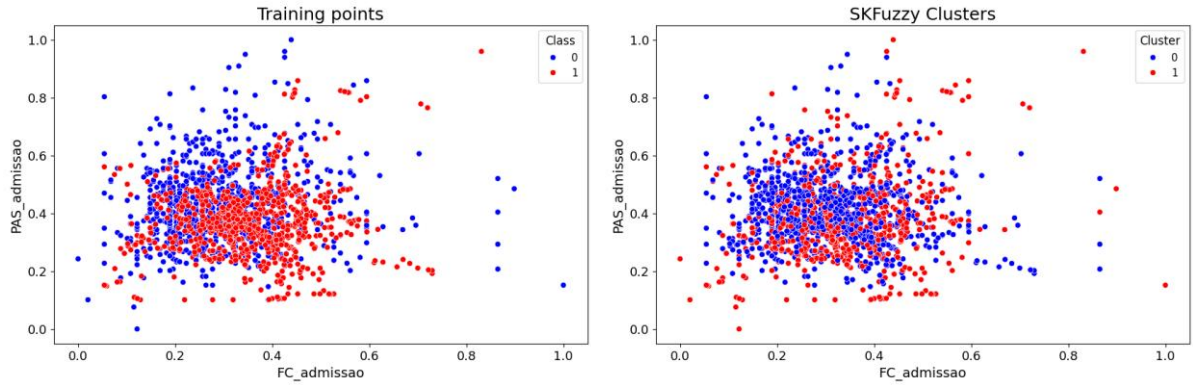


a) b)

Figure 5.15 – Results using σ_0: 0.139 and σ_1: 0.074: a) Number of correct/incorrect predictions for each reliability interval; b) Percentage number of correct/incorrect predictions for each reliability interval.

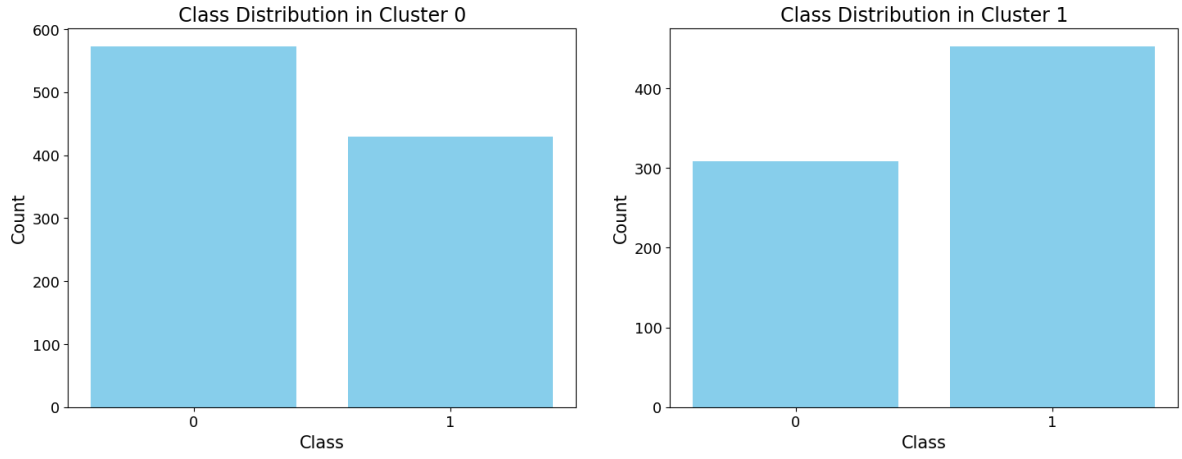
5.2.2 Clustering-based method

The application of the Clustering-based method to the CRA dataset was done with resource to the Fuzzy C-Means clustering algorithm, which used the following parameterization values, adjusted through experiments: $C = 2$; $m = 2$; *error* = 0.005; *maxiter* = 10000; *init* = 'None'. The Figure 5.16a displays the training points distribution by class, and Figure 5.16b the data distribution of the points by each cluster per two features: *FC_admissao* and *PAS_admissao*. The Figure 5.17 highlights the distribution of the classes across the two clusters formed using the training points.



a) b)

Figure 5.16 – Training points distribution: a) by class; b) by each cluster.



a) b)

Figure 5.17 – Number of training points per class across the clusters: a) cluster 0; b) cluster 1.

Both figures proof a high level of class blending in each of the two clusters. About 43% (429 data points) of the total data points belonging to cluster 0 are from class 1, and near 40% (309 data points) data points associated to cluster 1 are from class 0. This represents a significant mixing of data from both classes within each of the clusters. Figure 5.18 and Figure 5.19 present the results of the reliability measurement with the usage of these clusters.

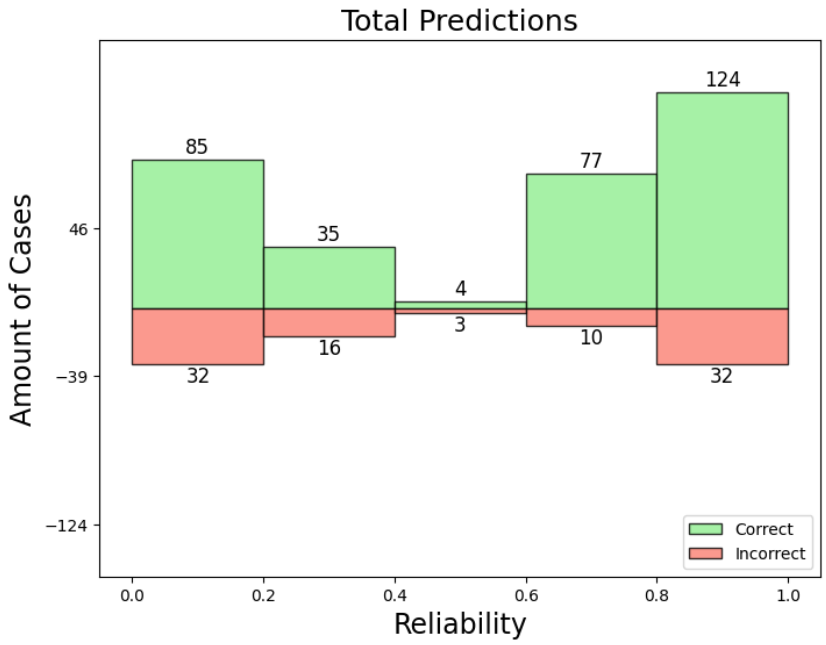


Figure 5.18 – Number of correct/incorrect predictions for each reliability interval using the CB method, on CRA dataset.

It is noticeable that no obvious relationship between the metric values and the ML model output seems evident. In fact, bimodal-like patterns appear to happen with two distinct peaks, at lower and higher reliabilities, with some variation in between. The incorrect predictions are the most important to minimize at higher reliabilities, however it results on an equal quantity of 32 cases, at the lowest reliabilities [0.0-0.2] and the highest reliabilities [0.8-1.0].

As speculated, Figure 5.19 confirms the lack of relation between reliability and prediction output. No trend is observed, and is very likely due to the multi-class clusters, not allowing a clear class assignment per cluster, resulting in an unclear reliability diagnosis on test points. Further experiments were directed in order to bypass the problem, such as other clustering algorithms, removing the majority class instances in the class-overlap area and dimensionality reductions (PCA), however none of these efforts returned successful results.

This absence of relation is validated by the Point Biserial Correlation which, when applied to the Clustering-based method using the CRA dataset, a coefficient value of 0.13 is obtained with a *p-value* = 0.007. A correlation of 0.13 in a range of [-1,1] suggests a weak relationship between reliability values and prediction outcomes, and the *p-value* indicates that the correlation is statistically significant, suggesting that the weak correlation observed is unlikely to have occurred by random chance alone.

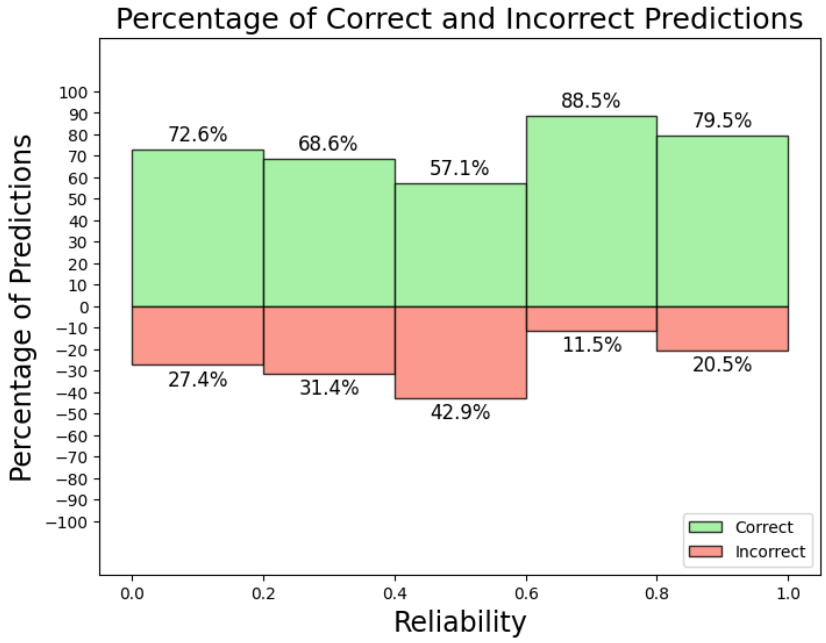


Figure 5.19 – Percentage number of correct/incorrect predictions for each reliability interval using the CB method, on CRA dataset.

5.2.3 Convex Hull-based method

The application of the Convex Hull-based method to the CRA dataset involves the tuning of parameters such as α, γ and β. Initially and according to what was done for the BMI dataset, the parameters α and γ were based on the average distance of the training points to the border of the convex hull, but were adjusted through successive experiments, by analyzing the final results of each experiment. The parameter β is a value close to zero, so that there isn't much tolerance for when the points are outside the convex hull. Figure 5.20 and Figure 5.21 represent the results when using these parameters tunings, generating α: 0.08, γ: 0.02 and β: 0.002.

Figure 5.20 represents the amount of correct and incorrect predictions corresponding to the predictions of the ML model *A*. The results show a clear increase of correct predictions, and a decrease of incorrectly classified predictions, with the reliability escalation. The majority of the correctly classified predictions (154) are located at the highest reliabilities (0.8-1.0), while the majority of the incorrectly classified predictions (43) are located at the lowest reliabilities (0.0-0.2).

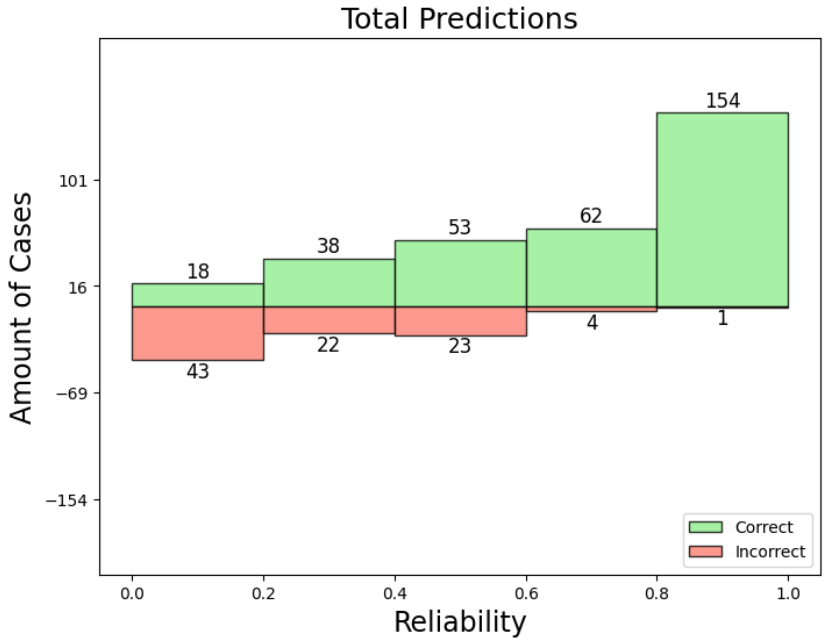


Figure 5.20 – Number of correct/incorrect predictions for each reliability interval using the CHB method, on CRA dataset; α: 0.08; γ: 0.02; β: 0.002.

Figure 5.21 shows the percentage of each prediction type on certainty intervals, enforcing the evolution of correct and incorrect predictions along the reliability scale. It's possible to perceive that the proportion of correctly classified predictions tends to outgrow the proportion of incorrectly classified predictions as the reliability increases. Even though the medial reliability scores (0.4-0.6) don't represent a near half proportion of correct and incorrect predictions, it's observable that the incorrect predictions are predominant at the lowest reliability scores (0.0-0.2).

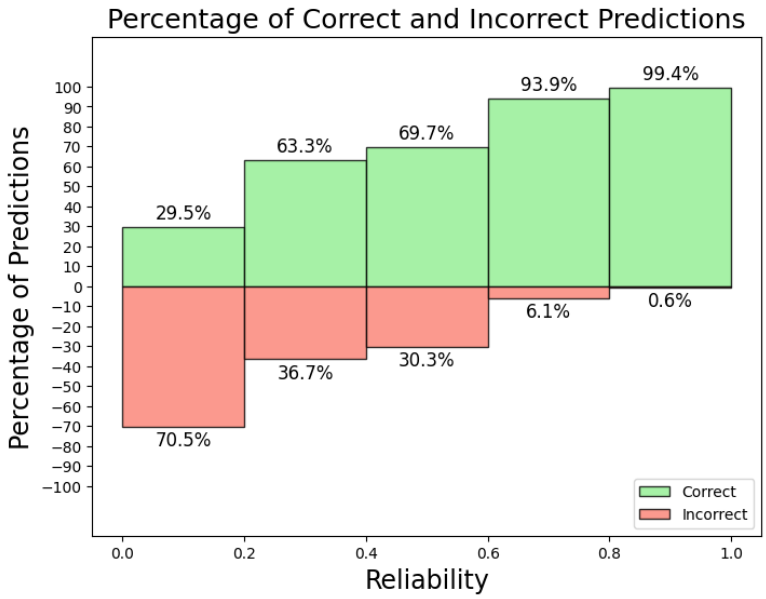


Figure 5.21 – Percentage number of correct/incorrect predictions for each reliability interval using the CHB method, on CRA dataset; α: 0.08; γ: 0.02; β: 0.002.

Applying the Point Biserial Correlation to the Convex Hull-based method using the CRA dataset, a coefficient value of 0.57 was obtained with a *p-value* ≈ 0 (1.24×10^{-37}). A correlation of 0.57 in a range of [-1,1] indicates a meaningful positive association between reliability values and prediction outcomes, and the extremely close to zero *p-value* value indicates that the correlation is highly statistically significant.

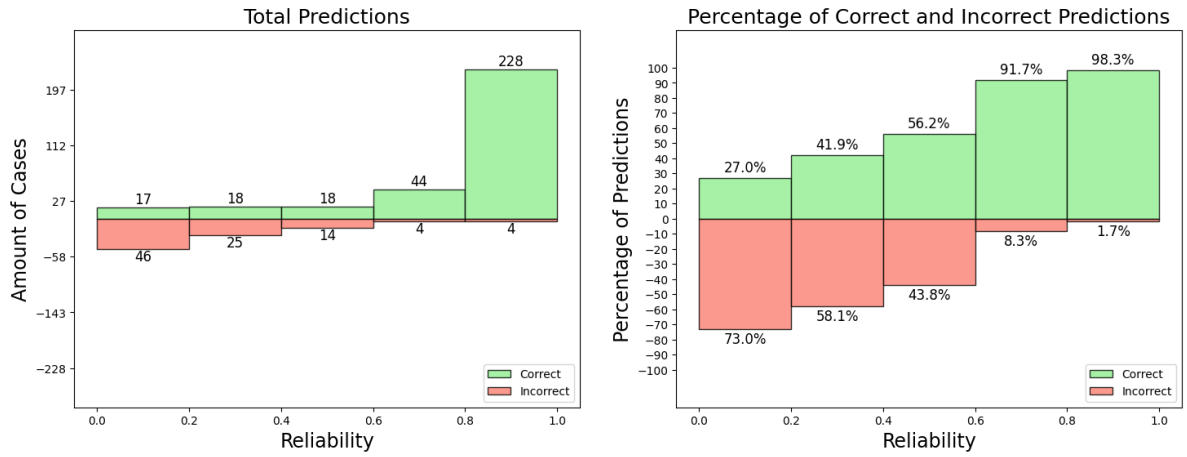
A further enhancement of the results based on tuning of the parameters was done by analyzing a scale of [0.01-1.0] on both parameters. This method was also done in order to develop an automatic evaluation of the parameters tuning process, instead of a manual trial and error process. Table 5.1 shows the Point Biserial Correlation coefficient values for each different combination of α and γ.

Table 5.1 – Point Biserial Correlation coefficient values.

α/γ	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	**0.1**
0.01	0.23	0.32	0.37	0.40	0.42	0.44	0.45	0.46	0.47	0.47
0.02	0.32	0.41	0.46	0.49	0.51	0.52	0.53	0.54	0.54	0.55
0.03	0.38	0.47	0.51	0.54	0.56	0.57	0.58	0.58	0.59	0.59
0.04	0.41	0.50	0.55	0.57	0.59	0.60	0.61	0.61	0.62	0.62
0.05	0.44	0.53	0.57	0.59	0.61	0.62	0.62	0.63	0.63	0.64
0.06	0.46	0.55	0.59	0.61	0.62	0.63	0.64	0.64	0.65	0.65
0.07	0.48	0.56	0.60	0.62	0.63	0.64	0.65	0.65	0.66	0.66
0.08	0.49	0.57	0.61	0.63	0.64	0.63	0.65	0.65	0.66	0.66
0.09	0.50	0.58	0.61	0.63	0.64	0.64	0.65	0.65	0.66	0.66
0.1	0.51	0.59	0.62	0.64	0.65	0.65	0.66	0.66	0.67	0.67

All the results were linked with an extremely small *p-value* of ≈ 0. Based on the results obtained, the best tuning of α and γ is 0.1 for both parameters, because it provides the highest coefficient value of 0.67.

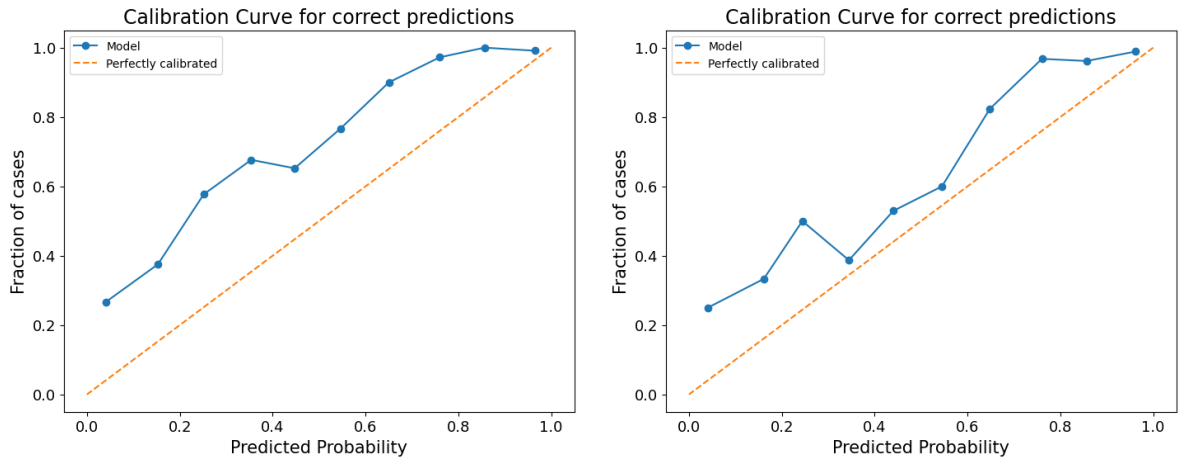
The application of the Convex Hull-based method to the CRA dataset was repeated with the new parameters values, which produced the results in Figure 5.22.



a) b)

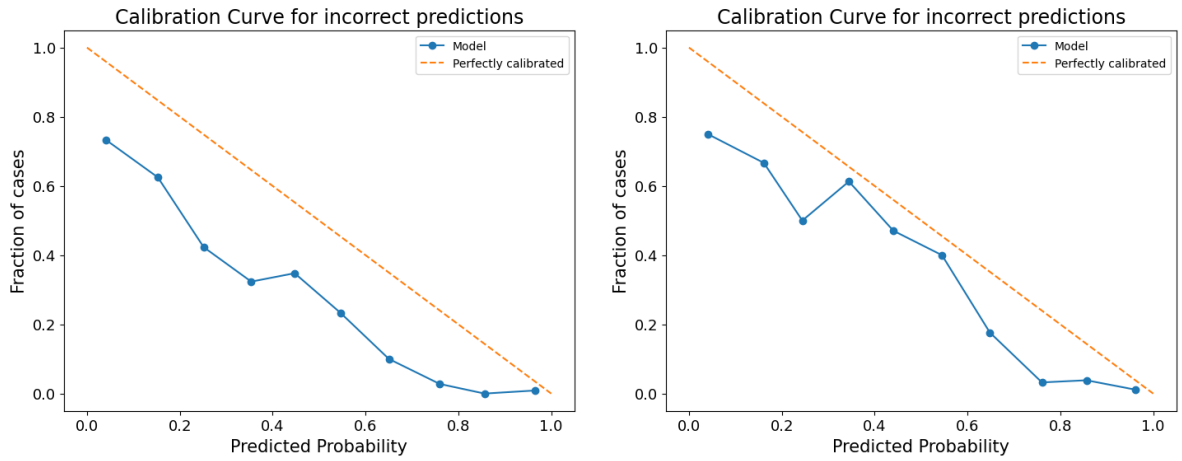
Figure 5.22 – Results using α: 0.1; γ: 0.1; β: 0.002: a) Number of correct/incorrect predictions for each reliability interval; b) Percentage number of correct/incorrect predictions for each reliability interval.

These new results show a slightly more evenly distribution across the reliability range, but still very similar to the original parameters tuning, offering an automated optimization parameters process. This conclusion is supported by the analysis of the calibration curve for the correct and incorrect predictions show in Figure 5.23 and Figure 5.24, respectively. The calibration curve provides a perspective of how close the results align with the ideal balanced output.



a) b)

Figure 5.23 – Calibration curve for correct predictions: a) model with α: 0.08; γ: 0.02; b) tuned model with α: 0.1; γ: 0.1.



a) b)

Figure 5.24– Calibration curve for incorrect predictions: a) model with α: 0.08; γ: 0.02; b) tuned model with α: 0.1; γ: 0.1.

The Figure 5.23 and Figure 5.24 show the fraction of correctly classified and incorrectly classified cases, respectively, per each reliability probability. A perfectly balanced model would follow the Perfectly calibrated line, having about 50/50 cases of correct and incorrect predictions on outputs of 50% reliability, and an absolute fraction of incorrect predictions for reliabilities near 0%, declining linearly as the reliability escalates. Even though the model with automated tuning process (Figure 5.23b and Figure 5.24b) is slightly closer to the ideal scenario, both models present similar positive results.

5.3 Discussion

Based on the previously presented results, some of the proposed methods seem to be more promising, while others aren't able to effectively quantify the reliability of the model, at least for the two datasets used in this study. The Data Proximity-based method is the method with most potential, followed by the Convex Hull-based method. And lastly the Clustering-based method, which seems to be unsuitable for the intended purposes.

The Data Proximity-based method successfully assigns the highest levels of reliability to correct predictions in majority, and the lowest levels of reliability to mainly incorrect predictions. Also, the progression of correct predictions through the reliability follows a gradual growth of cases, while the incorrect predictions decrease, depending on the defined σ values. This factor means the method has a degree of ambiguity to some data points, preferably allocating them to medium values than taking the risk to wrongfully assigning to a high level of reliability. These

observations happen in both the BMI dataset as well in the CRA dataset, strengthening the concept of being independent of the data used.

The Convex Hull-based method also effectively captures higher levels of reliability for correct predictions, and the opposite for incorrect predictions. However, the appliance on the BMI dataset couldn't ideally capture the desired evolution of correctly predicted cases through the reliability. These effects are probable to happen on datasets with a high overlap of classes, as seen in the BMI dataset, and when the parameters σ and γ aren't perfectly tuned for the case. Still, even though the overall increase of correct predictions wasn't captured, the positive class - on which the convex hull was based - was accurately assigned reliability levels, proving to be valuable on the analysis of specific classes, including FN and TP. To be reminded the BMI dataset is synthetic, and therefore, its data distribution is not to be taken into consideration as highly as the CRA, and other real-world datasets.

The Clustering-based method couldn't reproduce the desired outcomes. The reason for that, is the high blending of classes in clusters. Clustering is a technique that groups data points based on their characteristics, without prior knowledge of their true class label. To use clustering as a reliability method, the data would need to allow for a perfect or near-perfect class division, something that it's not possible on most scenarios, including the datasets used on this project. The different trends observed when applying the BMI and CRA datasets can be attributed to the distinct features in both datasets, which cause more distinguishable clusters. The BMI has less distinguishable clusters, resulting in closer cluster centers (Euclidean distance of 0.3), while the CRA dataset, shows more spread-out clusters with a distance of 0.9.

When applying the Data Proximity and Convex Hull-based methods to both datasets, the Point Biserial Correlation showed strong positive associations between reliability values and prediction outcomes, with a p-value close to zero, indicating high statistical significance. On the contrary, when applying the Clustering-based method, this correlation is weaker, highlighting the negative results previously recognized with the application of this method.

In summary, the results demonstrate that while the methods do not perform equally, the Data Proximity and Convex Hull-based approaches demonstrate strong potential for assessing pointwise reliability, effectively capturing the relationship between reliability scores and prediction outcomes (correct or incorrect). The Clustering-based approach, on the other hand, proved to be less effective due to class overlap within the clusters.

6 CONCLUSIONS

The use of ML and AI in healthcare has grown significantly in recent years, especially in cardiology, where cardiovascular diseases remain a leading cause of death worldwide. However, the lack of trustworthiness in such models presents challenges, since medical decisions affect the well-being of people. The objective of this work was to implement pointwise reliability metrics that can evaluate the predictions made by ML models, with primary focus on the clinical field, especially in predicting cardiovascular risk.

In this study, the reliability assessment of single ML predictions was carried out based on three different methods: (1) based on data proximity inspired by the previously published Intuitive Certainty Metric, (2) considering clustering techniques, and (3) based on convex hulls. Such reliability methods address the assessment of individual instances on a [0,1] degree, in addition to the common metrics used to estimate the overall performance of a ML model. The developed methods were first applied to a synthetic dataset and then to a real-world dataset collected in a hospital environment provided by CHUC. The performance of the proposed methods was evaluated by observing and quantifying the evolution of correct/incorrect predictions through the reliability values.

The method (1) was the most promising method overall for evaluating individual predictions, followed by method (3), which showed its best performance when applied to the real-world dataset. The method (2) was found to be unsuitable for the purpose intended. The successful application of the methods is directly related to the parameters tuning belonging to each, and the quality of the dataset in use. Aside from that, the proposed methods are model independent, because they don't rely on any specific ML model and only make usage of distance measures and data patterns, based on density principles.

Taking everything into account, this work seeks to contribute to the integration of reliability assessment in ML models. It is based on studies related to explainability and trust in ML, which supported the development of the proposed methods. These methods aim to evaluate pointwise reliability, particularly in the context of medical disease prevention, with a focus on cardiovascular risk assessment.

Future research should involve testing the proposed methods on additional datasets to validate their effectiveness. It should also place greater emphasis on determining optimal parameter values, potentially through automated approaches. Furthermore, improving the current methods and developing new ones for assessing the reliability of ML predictions is critical for the progress of this still relatively underexplored area.

To conclude, it is important to highlight that the results obtained in this work have led to two scientific publications, listed below:

- J. Sousa, T. Rocha, S. Paredes, J. Henriques, and L. Gonçalves, "Certainty measurement of ML model prediction outcomes in the healthcare context," in *Proc. 30th Portuguese Conference on Pattern Recognition (RECPAD 2024)*, Covilhã, Portugal, Oct. 2024.
- J. Sousa, L. Gonçalves, S. Paredes, T. Rocha, and J. Henriques, "Pointwise reliability metrics for machine learning model predictions: three different approaches," in *Proceedings of the 2024 IEEE International Conference on Data Mining (ICDM 2024)*, 2024.

REFERENCES

[1] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare" *IEEE* Access, vol. 8, pp. 107562-107582, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9112202

[2] S. M. McKinney *et al.*, "International evaluation of an AI system for breast cancer screening" *Nature*, vol. 577, no. 7788, pp. 89–94, Jan. 2020. [Online]. Available: https://doi.org/10.1038/s41586-019-1799-6.

[3] J. M. Stokes *et al.*, "A deep learning approach to antibiotic discovery" *Cell*, vol. 180, no. 4, pp. 688–702.e13, Feb. 2020. [Online]. Available: https://doi.org/10.1016/j.cell.2020.01.021.

[4] F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA, USA: Harvard Univ. Press, 2015.

[5] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017. [Online]. Available: https://arxiv.org/abs/1712.09923

[6] World Heart Federation, *World Heart Report 2023*, 2023. [Online]. Available: https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf

[7] M. Di Cesare, P. Perel, S. Taylor, C. Kabudula, H. Bixby, T. A. Gaziano, *et al.*, "The heart of the world" *Global Heart*, vol. 19, no. 1, 2024.

[8] NHS, "Cardiovascular disease". [Online]. Available: https://www.nhs.uk/conditions/cardiovascular-disease/ (accessed Feb. 10, 2025).

[9] F. Valente, J. Henriques, S. Paredes, T. Rocha, P. de Carvalho, and J. Morais, "A new approach for interpretability and reliability in clinical risk prediction: Acute coronary syndrome scenario" Artif. Intell. Med., vol. 117, p. 102113, 2021. [Online]. Available: https://doi.org/10.1016/j.artmed.2021.102113

[10] World Health Organization, "Cardiovascular diseases" *WHO*, 2024. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. (accessed Feb. 11, 2025).

[11] World Health Organization, "Cardiovascular diseases (CVDs)" *WHO*, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Feb. 11, 2025).

[12] S. S. Martin, A. W. Aday, Z. I. Almarzooq, C. A. Anderson, P. Arora, C. L. Avery, *et al.*, "2024 heart disease and stroke statistics: A report of US and global data from the American Heart Association" *Circulation*, vol. 149, no. 8,

pp. e347–e913, 2024. [Online]. Available: https://doi.org/10.1161/CIR.0000000000001209

[13] A. Goldstein, *Heart and Circulatory Disease Statistics 2021*, British Heart Foundation, United Kingdom, 2021. [Online]. Available: https://coilink.org/20.500.12592/vj1qzs

[14] R. Beaglehole and R. Bonita, "Global public health: A scorecard" *The Lancet*, vol. 372, no. 9654, pp. 1988–1996, 2008. [Online]. Available: https://doi.org/10.1016/S0140-6736(08)61558-5

[15] B. Dahlöf, "Cardiovascular disease risk factors: Epidemiology and risk assessment" *Am. J. Cardiol.*, vol. 105, no. 1, pp. 3A–9A, 2010. [Online]. Available: https://doi.org/10.1016/j.amjcard.2009.10.007

[16] B. A. Hemann, W. F. Bimson, and A. J. Taylor, "The Framingham Risk Score: An appraisal of its benefits and limitations" *Am. Heart Hosp. J.*, vol. 5, no. 2, pp. 91–96, 2007. [Online]. Available: https://doi.org/10.1111/j.1541-9215.2007.06350.x

[17] D. L. Bhatt, P. G. Steg, E. M. Ohman, A. T. Hirsch, Y. Ikeda, J. L. Mas, *et al.*, "International prevalence, recognition, and treatment of cardiovascular risk factors in outpatients with atherothrombosis" *JAMA*, vol. 295, no. 2, pp. 180–189, 2006. [Online]. Available: https://doi.org/10.1001/jama.295.2.180

[18] N. J. Bosomworth, "Practical use of the Framingham risk score in primary prevention: Canadian perspective," *Canadian Family Physician*, vol. 57, no. 4, pp. 417-423, 2011

[19] P. de Araújo Gonçalves, J. Ferreira, C. Aguiar, and R. Seabra-Gomes, "TIMI, PURSUIT, and GRACE risk scores: sustained prognostic value and interaction with revascularization in NSTE-ACS" *Eur. Heart J.*, vol. 26, no. 9, pp. 865–872, May 2005. [Online]. Available: https://doi.org/10.1093/eurheartj/ehi187

[20] K. A. Fox, O. H. Dabbous, R. J. Goldberg, K. S. Pieper, K. A. Eagle, F. Van de Werf, *et al.*, "Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE)" *BMJ*, vol. 333, no. 7578, p. 1091, 2006. [Online]. Available: https://doi.org/10.1136/bmj.38985.646481.55

[21] M. Roseiro, J. Henriques, S. Paredes, T. Rocha, and J. Sousa, "An interpretable machine learning approach to estimate the influence of inflammation biomarkers on cardiovascular risk assessment" *Comput. Methods Programs Biomed.*, vol. 230, p. 107347, 2023. [Online]. Available: https://doi.org/10.1016/j.cmpb.2023.107347

[22] B. Bawamia, R. Mehran, W. Qiu, and V. Kunadian, "Risk scores in acute coronary syndrome and percutaneous coronary intervention: A review" *Am.*

Heart J., vol. 165, no. 4, pp. 441–450, 2013. [Online]. Available: https://doi.org/10.1016/j.ahj.2012.12.020

[23] J. McCarthy, "What is artificial intelligence" *Stanford University*, 2007. [Online]. Available: http://jmc.stanford.edu/articles/whatisai.html. (accessed: Feb. 12, 2025)

[24] Markets and Markets, "Artificial Intelligence (AI) Market" *Markets and Markets*, 2023. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-market-74851580.html. (accessed: Feb. 13, 2025)

[25] A. Balodi, *Application of Introduction Artificial Intelligence Machine Learning in Real Life*, 2020. [Online]. Available: https://www.researchgate.net/publication/340684782_Application_of_Introduction_Artificial_Intelligence_Machine_Learning_in_Real_Life

[26] Z. H. Zhou, *Machine Learning*, Springer Nature, 2021.

[27] G. Rebala, A. Ravi, and S. Churiwala, *An Introduction to Machine Learning*, Springer, 2019.

[28] V. Nasteski, "An overview of the supervised machine learning methods" *Horizons B*, vol. 4, pp. 51–62, 2017. [Online]. Available: https://doi.org/10.20544/HORIZONS.B.04.1.17.P05.

[29] B. Mahesh and B. Batta, "Machine Learning Algorithms - A Review" *Int. J. Sci. Res. (IJSR)*, vol. 9, no. 1, pp. 1–5, 2019. [Online]. Available: https://doi.org/10.21275/ART20203995.

[30] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1702.08608

[31] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2021. [Online]. Available: https://doi.org/10.3390/e23010018

[32] R. Dwivedi *et al.*, "Explainable AI (XAI): Core ideas, techniques, and solutions" *ACM Comput. Surv.*, vol. 55, no. 9, Art. 194, pp. 1–33, Sep. 2023. [Online]. Available: https://doi.org/10.1145/3561048.

[33] L. J. Cronbach, "Test reliability: Its meaning and determination" *Psychometrika*, vol. 12, no. 1, pp. 1–16, 1947. [Online]. Available: https://doi.org/10.1007/BF02289289

[34] R. J. Mislevy, "Can there be reliability without 'reliability'?" *J. Educ. Behav. Stat.*, vol. 29, no. 2, pp. 241–244, 2004. [Online]. Available: https://doi.org/10.3102/10769986029002241

[35] G. Nicora, M. Rios, A. Abu-Hanna, and R. Bellazzi, "Evaluating pointwise reliability of machine learning prediction" *J. Biomed. Informatics*, vol. 127, p. 103996, 2022. [Online]. Available: https://doi.org/10.1016/j.jbi.2022.103996

[36] S. Saria and A. Subbaswamy, "Tutorial: Safe and reliable machine learning" *arXiv preprint arXiv:1904.07204*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1904.07204

[37] R. Xu and D. Wunsch, *Clustering*, John Wiley & Sons, 2008

[38] F.-J. Yang, "An implementation of Naive Bayes classifier" in *Proc. 2018 Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Las Vegas, NV, USA, 2018, pp. 301–306. [Online]. Available: https://doi.org/10.1109/CSCI46756.2018.00065

[39] M. Bobokulova, "Uncertainty in the Heisenberg uncertainty principle" *Akademicheskie Issledovaniya v Sovremennoi Nauke*, vol. 3, no. 2, pp. 80–96, 2024

[40] J. J. Zhang, J. N. Chen, D. Y. Meng, *et al.*, "Exploring the uncertainty principle in neural networks through binary classification" *Sci. Rep.*, vol. 14, p. 28402, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-79028-4

[41] K. Goher, A. Al-Ashaab, S. Sarfraz, and E. Shehab, "An uncertainty management framework to support model-based definition and enterprise" *Computers in Industry*, vol. 150, p. 103944, 2023. [Online]. Available: https://doi.org/10.1016/j.compind.2023.103944

[42] K. Zou, Z. Chen, X. Yuan, X. Shen, M. Wang, and H. Fu, "A review of uncertainty estimation and its application in medical imaging" *Meta-Radiology*, vol. 1, no. 1, p. 100003, 2023. [Online]. Available: https://doi.org/10.1016/j.metrad.2023.100003

[43] M. R. Smith *et al.*, "SAGE intrusion detection system: Sensitivity analysis guided explainability for machine learning" 2021. [Online]. Available: https://doi.org/10.2172/1820253

[44] C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high-dimensional space," *Lecture Notes in Comput. Sci.*, vol. 1973, pp. 420–434, 2001.

[45] A. Ghodsi, "Dimensionality reduction: A short tutorial" Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada, vol. 37, no. 38, 2006.

[46] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson Education India, 2016.

[47] J. C. Gower, "A general coefficient of similarity and some of its properties" *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971. [Online]. Available: https://doi.org/10.2307/2528823

[48] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine Learning Proceedings 1994*, W. W. Cohen and

H. Hirsh, Eds. San Francisco, CA, USA: Morgan Kaufmann, 1994, pp. 148–156. [Online]. Available: https://doi.org/10.1016/B978-1-55860-335-6.50026-X

[49] G. Shafer and V. Vovk, "A tutorial on conformal prediction" *Journal of Machine Learning Research*, vol. 9, no. 3, 2008

[50] C. Saunders, A. Gammerman, and V. Vovk, "Transduction with confidence and credibility" in *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI '99)*, 1999, pp. 722–726. [Online]. Available: https://eprints.soton.ac.uk/258961/

[51] H. Papadopoulos, V. Vovk, and A. Gammerman, "Conformal prediction with neural networks" *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, Patras, Greece, 2007, pp. 388-395. [Online]. Available: https://doi.org/10.1109/ICTAI.2007.47

[52] B. Hanczar and E. R. Dougherty, "Classification with reject option in gene expression data" *Bioinformatics*, vol. 24, no. 17, pp. 1889–1895, Sep. 2008. [Online]. Available: https://doi.org/10.1093/bioinformatics/btn349

[53] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection" *arXiv preprint arXiv:2003.10769*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2003.10769

[54] A. Shaik *et al.*, "A staged approach using machine learning and uncertainty quantification to predict the risk of hip fracture" *Bone Reports*, vol. 22, p. 101805, 2024. [Online]. Available: https://doi.org/10.1016/j.bonr.2024.101805

[55] G. Nicora and R. Bellazzi, "A Reliable Machine Learning Approach applied to Single-Cell Classification in Acute Myeloid Leukemia" *AMIA Annu Symp Proc.*, vol. 2020, pp. 925–932, Jan. 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33936468

[56] J. Waa, J. Diggelen, M. Neerincx, and S. Raaijmakers, "ICM: An Intuitive Model Independent and Accurate Certainty Measure for Machine Learning" in *Proc. 10th Int. Conf. Agents Artif. Intell. (ICAART)*, 2018, vol. 2, pp. 314–321. [Online]. Available: https://doi.org/10.5220/0006542603140321

[57] J. Henriques, T. Rocha, S. Paredes, P. Gil, J. Loureiro, and L. Petrella, "Pointwise Reliability of Machine Learning Models: Application to Cardiovascular Risk Assessment" in *9th European Medical and Biological Engineering Conference (EMBEC 2024)*, T. Jarm, R. Šmerc, and S. Mahnič-Kalamiza, Eds. Cham, Switzerland: Springer, 2024, vol. 113, pp. 23. [Online]. Available: https://doi.org/10.1007/978-3-031-61628-0_23

[58] A. Jan and C. Weir, "BMI Classification Percentile and Cut Off Points", 2019.

[59] N. Chawla, "Data Mining for Imbalanced Datasets: An Overview" 2005. [Online]. Available: https://doi.org/10.1007/0-387-25465-X_40

[60] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project" *PLoS One*, vol. 12, no. 7, p. e0179805, 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0179805

[61] C. Bustamante, L. Garrido, and R. Soto, "Comparing Fuzzy Naive Bayes and Gaussian Naive Bayes for Decision Making in RoboCup 3D" in *MICAI 2006: Advances in Artificial Intelligence*, A. Gelbukh and C. A. Reyes-Garcia, Eds., vol. 4293, *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, 2006. [Online]. Available: https://doi.org/10.1007/11925231_23

[62] K. W. Lau and Q. H. Wu, "Online training of support vector classifier" *Pattern Recognition*, vol. 36, no. 8, pp. 1913–1920, 2003. [Online]. Available: https://doi.org/10.1016/S0031-3203(03)00038-4

[63] Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). "Density-based clustering." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3), 231-240. [Online]. Available: https://doi.org/10.1002/widm.30

[64] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls" *ACM Transactions on Mathematical Software*, vol. 22, no. 4, pp. 469–483, Dec. 1996. [Online]. Available: https://doi.org/10.1145/235815.235821

[65] D. Avis and D. Bremner, "How good are convex hull algorithms?" in *Proceedings of the 11th Annual Symposium on Computational Geometry*, Sept. 1995, pp. 20–28

[66] D. Kornbrot, "Point biserial correlation" *Wiley StatsRef: Statistics Reference Online*, 2014. [Online]. Available: https://doi.org/10.1002/9781118445112.stat07898

Instituto Superior de Engenharia
Politécnico de Coimbra