NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

ISABEL DE ALMEIDA CURIOSO

Bachelor of Science in Biomedical Engineering

DELIVERING RELIABLE AI TO CLINICAL CONTEXTS: ADDRESSING THE CHALLENGE OF MISSING DATA

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
December, 2022

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

DELIVERING RELIABLE AI TO CLINICAL CONTEXTS: ADDRESSING THE CHALLENGE OF MISSING DATA

ISABEL DE ALMEIDA CURIOSO

Bachelor of Science in Biomedical Engineering

Adviser: Dr. Hugo Filipe Silveira Gamboa
Associate Professor with Aggregation, NOVA University of Lisbon

Examination Committee

Chair: Dr. Ricardo Nuno Pereira Verga e Afonso Vigário
Associate Professor with Aggregation, NOVA University of Lisbon

Rapporteur: Dr. Hui Liu
Researcher, University of Bremen

Adviser: Dr. Hugo Filipe Silveira Gamboa
Associate Professor with Aggregation, NOVA University of Lisbon

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
December, 2022

Delivering Reliable AI to Clinical Contexts: Addressing the Challenge of Missing Data

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.9.15) [1].

Acknowledgements

There were many remarkable people who shaped my academic journey, and to whom I would like to express my gratitude. Know that these words will never be enough to fully convey the immensity of my appreciation.

First of all, to my advisor, Professor Hugo Gamboa, for the vote of confidence which allowed me the opportunity to explore this fascinating subject, and for encouraging me to do my best.

To Fraunhofer AICOS, particularly to the Lisbon team, for the warm welcome, amazing work environment, moments of leisure and words of advice. I would like to specially thank Ricardo and Bruno, who answered all my questions and were available at any time, especially in the final stage of the dissertation writing, when no day passed without me having new concerns. I was extremely lucky for their motivation, relentless support, and immeasurable patience.

To the amazing friends that FCT has given me, with whom I shared unforgettable moments and memories that I will cherish forever. I am truly grateful to have met you.

To Zé, who has stuck by my side for the last 7 years, makes me see the brighter side of things, and supports me unconditionally.

To my mother and my father, who always cared about me and told me that they would be proud of me no matter what. To my sister, my person, for always lifting me up and showing me that every problem had an answer. To my brother, for checking in on me and making me laugh when I needed it most. All my achievements were and will always be due to my wonderful and unique family.

Abstract

Clinical data are essential in the medical domain, ensuring quality of care and improving decision-making. However, their heterogeneous and incomplete nature leads to an ubiquity of data quality problems, particularly missing values. Inevitable challenges arise in delivering reliable Decision Support Systems (DSSs), as missing data yield negative effects on the learning process of Machine Learning models. The interest in developing missing value imputation strategies has been growing, in an endeavour to overcome this issue.

This dissertation aimed to study missing data and their relationships with observed values, and to later employ that information in a technique that addresses the predicaments posed by incomplete datasets in real-world scenarios. Moreover, the concept of correlation was explored within the context of missing value imputation, a promising but rather overlooked approach in biomedical research.

First, a comprehensive correlational study was performed, which considered key aspects from missing data analysis. Afterwards, the gathered knowledge was leveraged to create three novel correlation-based imputation techniques. These were not only validated on datasets with a controlled and synthetic missingness, but also on real-world medical datasets. Their performance was evaluated against competing imputation methods, both traditional and state-of-the-art.

The contributions of this dissertation encompass a systematic view of theoretical concepts regarding the analysis and handling of missing values. Additionally, an extensive literature review concerning missing data imputation was conducted, which comprised a comparative study of ten methods under diverse missingness conditions. The proposed techniques exhibited similar results when compared to their competitors, sometimes even superior in terms of imputation precision and classification performance, evaluated through the Mean Absolute Error and the Area Under the Receiver Operating Characteristic curve, respectively. Therefore, this dissertation corroborates the potential of correlation to improve the robustness of DSSs to missing values, and provides answers to current flaws shared by correlation-based imputation strategies in real-world medical problems.

Keywords: Missing Data, Missing Data Imputation, Correlation, Machine Learning, Decision Support System

Resumo

Dados clínicos são essenciais para assegurar cuidados médicos de qualidade e melhorar a tomada de decisões. Contudo, a sua natureza heterogénea e incompleta cria uma ubiquidade de problemas de qualidade, nomeadamente pela existência de valores em falta. Esta condição origina desafios inevitáveis para a disponibilização de Sistemas de Apoio à Decisão (SADs) fiáveis, já que dados em falta acarretam efeitos negativos no treino de modelos de Aprendizagem Automática. O interesse no desenvolvimento de estratégias de imputação de valores em falta tem vindo a crescer, num esforço para superar esta adversidade.

Esta dissertação visou estudar o problema dos dados em falta através das relações que estes apresentam com os valores observados. Esta informação foi depois utilizada no desenvolvimento de técnicas para colmatar os problemas impostos por dados incompletos em cenários reais. Ademais, o conceito de correlação foi explorado no contexto da imputação de valores em falta, já que, apesar de promissor, tem vindo a ser negligenciado em investigação biomédica.

Em primeiro lugar, foi realizado um estudo correlacional abrangente que contemplou aspetos fundamentais da análise de dados em falta. Posteriormente, o conhecimento recolhido foi aplicado na criação de três novas técnicas de imputação baseadas na correlação. Estas foram validadas não só em conjuntos de dados com incompletude controlada e sintética, mas também em conjuntos de dados médicos reais. O seu desempenho foi avaliado e comparado a métodos de imputação tanto tradicionais como de estado-de-arte.

As contribuições desta dissertação passam pela sistematização de conceitos teóricos relativos à análise e tratamento de dados em falta. Adicionalmente, realizou-se uma extensa revisão da literatura referente à imputação de dados, que compreendeu um estudo comparativo de dez métodos sob diversas condições de incompletude. As técnicas propostas exibiram resultados semelhantes aos dos restantes métodos, por vezes até superiores em termos de precisão da imputação e de performance da classificação. Assim, esta dissertação corrobora o potencial da utilização da correlação na melhoria da robustez de SADs a dados em falta, e fornece respostas a algumas das atuais falhas partilhadas por estratégias de imputação baseadas em correlação quando aplicadas a casos médicos reais.

Palavras-chave: Dados em Falta, Imputação de Dados em Falta, Correlação, Aprendizagem Automática, Sistema de Apoio à Decisão

Contents

List of Figures

List of Tables

Abbreviations

AI Artificial Intelligence *(pp. 1, 2, 13, 37, 39, 94–96)*

AUROC Area Under the Receiver Operating Characteristic *(pp. ix–xi, 21, 22, 31, 62, 63, 87–93, 95, 114, 126, 127, 130)*

CCMVI Class Center based Missing Value Imputation *(pp. 33, 34, 36)*

CMIM Correlation Maximization-based Imputation Methods *(pp. 33, 35, 54, 55, 82, 83, 86, 88, 90, 92)*

CNNI Convolutional Neural Network Imputation *(pp. 35, 37)*

CoHiKNN Correlation-based Hierarchical K-Nearest Neighbors *(pp. 33, 36, 54, 55, 82, 83, 86, 88, 90, 92)*

CRAN Comprehensive R Archive Network *(pp. xi, 50, 55, 130)*

CWKNNI Correlation Weighted K-Nearest Neighbour Imputation *(pp. viii, 55–59, 82, 83, 85–88, 90–93, 95, 96, 127, 128)*

CWRI Correlation Weighted Regression Imputation *(pp. viii, 55, 58, 60, 61, 83, 87–90, 92, 93, 95, 127)*

DL Deep Learning *(pp. 35, 37–39)*

DSS Decision Support System *(pp. 1, 2, 37, 39, 89, 94–96)*

DT Decision Tree *(pp. 13, 17, 33, 35, 38)*

EACTS European Association for Cardio-Thoracic Surgery *(p. 46)*

EHR Electronic Health Record *(pp. 1, 5, 6)*

EM Expectation Maximisation *(pp. 30, 35, 37)*

EMR Electronic Medical Record *(p. 5)*

FCMI Feature Correlation based Missing Data Imputation *(pp. 32, 35)*

FN False Negative *(pp. 19, 20)*

FP False Positive *(pp. 19–21)*

FPR False Positive Rate *(pp. 21, 22)*

ICU Intensive Care Unit *(p. 46)*

KNN K-Nearest Neighbour *(pp. 14, 28, 29, 32–39, 54, 55, 58, 82, 83, 88, 90, 92)*

KNNSCI K-Nearest Neighbours Selected by Correlation Imputation *(pp. viii, 55, 57–59, 82, 83, 85–90, 92, 93, 95, 96, 127, 128)*

MAE Mean Absolute Error *(pp. ix, x, 3, 31, 63, 67, 68, 70, 72, 77, 82–87, 89, 95, 125)*

MAR Missing at Random *(pp. viii, ix, 23–27, 29, 30, 32, 33, 35–37, 41–44, 50, 52, 53, 67, 68, 70–72, 74–77, 79–81, 83–86, 95, 117, 120, 123)*

MCAR Missing Completely at Random *(pp. viii, ix, 23–28, 30, 33–37, 41–44, 50, 67–73, 76–80, 83–86, 88, 90, 92, 95, 118, 119, 122)*

MI Multiple Imputation *(pp. 28, 29, 32–38)*

MICE Multivariate Imputation by Chained Equations *(pp. 29, 34, 37, 38, 54, 55, 82, 83, 88, 90, 92, 93, 127, 128, 130)*

MIMS Monitor-Independent Movement Summary *(p. 110)*

ML Machine Learning *(pp. viii–x, 1, 3, 5, 8, 13–18, 21, 31, 37–39, 48, 54, 61–63, 87, 89, 91, 95, 96, 126)*

MNAR Missing not at Random *(pp. viii, ix, 23–27, 29, 30, 36, 41–44, 50, 67–72, 74–77, 79–88, 90, 92, 95, 121, 124)*

MR Missing Rate *(pp. viii–x, 2, 3, 23, 27, 29, 33–37, 40, 41, 45–48, 50–52, 56, 59, 67–83, 85–92, 95, 109–113, 117–124)*

NB Naive Bayes *(pp. x, 14, 17, 18, 61–63, 87, 91–93, 126–128, 130)*

NHANES National Health and Nutrition Examination Survey *(p. 44)*

NMVI Nullify the Missing Values before Imputation *(pp. 34, 36, 54, 55, 82, 83, 88, 90, 92, 93, 127)*

PHR Personal Health Record *(p. 5)*

QUIP Quality Improvement Programme *(p. 46)*

RF Random Forest *(pp. x, 13, 17, 18, 33, 61–63, 87–89, 93, 126–128, 130)*

RMSE Root Mean Squared Error *(p. 30)*

ROC Receiver Operating Characteristic *(pp. viii, 21, 22)*

RQ Research Question *(pp. 2–4, 48, 64, 95)*

SBP Systolic Blood Pressure *(pp. x, 24–26)*

SICE Single Center Imputation from Multiple Chained Equation *(pp. 34, 37)*

SL Supervised Learning *(pp. 13, 14, 34)*

SPECT Single Proton Emission Computed Tomography *(pp. vi, viii–x, 40–42, 51, 66–68, 71, 72, 74, 75, 77–79, 85, 86, 89, 91, 116, 117, 122, 124, 125)*

SVM Support Vector Machine *(pp. viii, x, 14, 17, 34–36, 38, 39, 61–63, 87, 89, 90, 93, 126–128, 130)*

TN True Negative *(pp. 19–21)*

TNR True Negative Rate *(p. 21)*

TP True Positive *(pp. 19, 20)*

TPR True Positive Rate *(pp. 20–22)*

UCI University of California, Irvine *(pp. vi, x, 40, 41, 43, 48–54, 62, 63, 82, 89)*

Chapter 1

Introduction

1.1 Context and Motivation

A growing and ageing population entails a broad amount of clinical information that needs to be organised and easily accessed by professionals. Since the analysis of paper-based data is time-consuming and far from the ideal solution, digitalization has emerged as a vital process towards the optimisation of health care management, accompanied by an arising of useful tools such as Electronic Health Records (EHRs) [2].

An EHR contains thorough clinical information, and can thus facilitate knowledge acquisition. However, in most cases, the establishment of relationships between data in order to formulate a medical diagnosis still relies solely on the physician. Artificial Intelligence (AI), and specifically Machine Learning (ML), is suitable for discovering patterns in vast datasets, an ability that could benefit clinical decision-making and support health care decisions [3]. Therefore, AI-based Decision Support Systems (DSSs) will become an important instrument to assist professionals, leveraging all available information from the patient's journey.

Moreover, EHRs mirror the heterogeneous nature of clinical data, often collected through different procedures and stored in distinct formats. Unfortunately, with this variability also comes inconsistency. In fact, the majority of real-world datasets are incomplete, which yields deleterious effects on the learning process of a ML model [4].

A reliable AI-based DSS must be able to cope with missing values since its performance may influence clinical decision-making [5]. This concern, along with the ubiquity of missing data in any real-world database, prompted a growing interest in developing strategies that address this challenge, particularly missing value imputation techniques.

Nevertheless, state-of-the-art imputation approaches still face limitations that ought to be overcome, such as the requirement for the performance of a complete-case analysis.

Besides, a myriad of imputation methods is only validated on datasets with a controlled and synthetic missingness, which does not fully reflect the entropy of a real-world scenario.

In recent years, some authors have proposed techniques that account for correlation when imputing missing values, stating that such choice is beneficial [6], [7]. Correlation is a measure of association between two variables, and thus may provide helpful information to the prediction of missing values.

The concept of correlation gains relevance in clinical datasets, where distinct features frequently are different manifestations of the same physiological event or medical condition, consequently exhibiting a significant level of dependency. Although promising, the exploration of correlation within the context of missing value imputation is still scarce in biomedical research.

This dissertation does not merely analyse the correlation between values, but rather carries out a comprehensive correlational study to create novel imputation methods. The proposed strategies aim to improve the reliability of AI-based DSSs and to tackle the current flaws shared by correlation-based imputation techniques.

1.2 Objectives

The main objective of this dissertation is to address the challenges posed by missing inputs on real-world medical datasets, thus enhancing the robustness of AI-based DSS. The outlined approach was based on the exploration of probabilistic dependencies between attributes, more specifically correlation. The prospective benefits of considering correlation when imputing missing values in a real-world setting were assessed through an analysis of several existent imputation methods, which inspired the proposal of three novel imputation techniques.

The first stage of this work included a correlational study, i.e. an exploratory investigation of the associations between variables, conducted separately on every missingness source mechanism. A missingness mechanism describes the relationships between measured data and missingness, although not offering a causal explanation for the incompleteness.

Within this stage, missing values were injected into three complete datasets under all missingness mechanisms individually, with three different Missing Rates (MRs). Through the results of this stage, we aim to answer the following questions:

1. **Does a correlational study that considers missingness patterns provide useful insights to address missing data?**

    - **Research Question 1.1 -** Can correlation be used to distinguish between missingness mechanisms?
    - **Research Question 1.2 -** Does the Missing Rate affect the measured correlations?

In order to approach the main Research Question (RQ), the traditional correlation measurements between the values of two variables were extended. In fact, the correlation between the values of one variable and the missingness pattern of every other, as well as the correlation between the missingness patterns of two attributes were also analysed.

Furthermore, a study was conducted on how these types of correlation captured the relationships associated with each missingness mechanism. Distinctive traits were sought, therefore addressing **RQ 1.1**.

As for **RQ 1.2**, correlation matrices computed from datasets with different and known MRs were compared. The disparities were evaluated in terms of Mean Absolute Error (MAE) and standard deviation.

The second stage of this dissertation focused on evaluating missing data imputation procedures and their impact on the performance of ML models. In addition to the datasets with synthetically injected missing values, two real-world medical databases were used, enabling the performance evaluation to span across actual case studies. Three innovative imputation methods were developed and compared against existent techniques, both traditional and state-of-the-art. This second and last stage was designed to answer the question below:

2. **Can correlation be leveraged to create more robust and reliable missing data imputation methods?**

    - **Research Question 2.1 -** Does accounting for correlation enhance the imputation quality in different missingness mechanisms?
    - **Research Question 2.2 -** In real-world medical problems, does imputation based on correlation yield better classification results?
    - **Research Question 2.3 -** Does a more precise imputation improve the performance of a ML classifier?

The datasets with synthetically injected missing elements permit a comparison between original values and the estimations provided by every imputation method considered. An evaluation of the imputation quality was conducted for each missingness mechanism using the MAE, thus covering **RQ 2.1**.

Afterwards, the effect of the imputation procedure on the performance of ML classifiers was assessed. Apart from the aforementioned datasets, this evaluation also encompassed two real-world medical datasets. The selected imputation techniques were applied to these databases, and the imputed datasets were given as inputs to three ML models. In order to tackle **RQ 2.2**, the classification results yielded by each differently imputed dataset were analysed.

Lastly, in regards to **RQ 2.3**, the classification performance of models trained upon imputed datasets was compared against the performance of the model trained upon the original dataset. Since the original dataset has, by definition, a perfect imputation quality, an individual comparison of its performance with that of each imputed dataset addresses

RQ 2.3. Due to the need for an originally complete dataset, this assessment can not be accomplished on real-world databases.

1.3 Document Structure

This document is divided into seven Chapters. The first and current Chapter introduced the dissertation, discussing its motivation and main objectives. The second Chapter covers the fundamental theoretical concepts for the development of this dissertation. The third Chapter contemplates a literature review on missing value imputation, focusing on state-of-the-art techniques which exploit correlations between attributes. The fourth Chapter describes the five datasets used throughout this dissertation, detailing the types of variables and the distribution of missing values within every dataset. The fifth Chapter provides a comprehensive description of the methodologies adopted to fulfil the objectives of this work. The sixth Chapter presents and critically analyses the various results obtained within this project. Lastly, the seventh Chapter concludes the dissertation by demonstrating the main findings and contributions, as well as limitations and future work.

Chapter 2

Theoretical Concepts

This chapter covers several theoretical concepts essential for the development of this work, and is divided into four Sections. The first section provides a brief introduction to clinical data, focusing especially on EHRs. In the second section, the concept of correlation is discussed and several coefficients that measure the strength of this statistical association between two variables are presented. A proper understanding of these concepts is essential to study and leverage the relations between medical attributes. The third section contains an overview of the fundamental ML concepts. Lastly, the fourth section addresses missing data, particularly the missingness mechanisms and missing value handling approaches.

2.1 Electronic Health Records

In biomedical research, specially within the scope of clinically oriented investigation, it is common to come across complex and large-scale datasets, namely the standardised data structures in health care: EHRs, Electronic Medical Records (EMRs) and Personal Health Records (PHRs) [8]. Although used interchangeably, these three terms are not necessarily equal, as the databases may encompass different information.

EMRs are fundamentally a repository of legal records created by health care providers, e.g. clinical codes for billing purposes, electronic prescriptions, disease management protocols, and others [9]. PHRs are separate from EMRs and comprise clinically relevant information concerning a particular patient (the user), such as their allergies, medications, laboratory data and radiology reports [2]. Lastly, a patient's EHR integrate longitudinal data gathered during clinical care encounters throughout their pathway, including demographics, medications, vital signs, laboratory data, radiology reports, electronic prescriptions, and scanned documents [9]. This section will focus on the clinical data provided by EHRs.

EHRs contain thorough medical information produced by multiple sources, which is stored either in a structured or an unstructured format. The structured data may be divided into: numerical data, which includes laboratory test results and vital sign measurements; coded data, e.g. procedure and diagnosis codes; categorical data, such as medication records; and demographic information [10]. The unstructured data includes radiology reports, medical images, clinical notes, progress notes, discharge summaries and detailed descriptions of anything the physician considers helpful for diagnosis and treatment [8], [10].

Although the information is collected through different procedures and stored in distinct formats, it can occasionally explain the same physiological event or medical condition. For instance, Fjell et al. [11] observed that different types of cerebrospinal fluid biomarkers yielded complementary information. Furthermore, Sylvestre et al. [12] showed that the detection of hyperkalemia was significantly improved if knowledge from drug prescriptions and laboratory test results was combined. One of the greatest contributions of EHRs for biomedical investigation is that they facilitate the search and study of these associations amongst the data.

However, there are challenges associated with the use of EHRs for research. Denaxas and Morley [8] presented some examples: substantially missing or incomplete datasets, data collected at irregular time-points, and the integration of information from multiple sources. Given the key role that clinical data play in the improvement of medical decision-making, it is essential to address these issues and to ensure data quality.

2.2 Probabilistic Dependency between Variables

According to Pillai and Leong [13], probabilistic dependencies between variables can take two forms: mutual and non-directional, such as a correlation; or conditional and directional, e.g. conditional dependence / independence and causation.

As aforementioned, distinct features in a clinical dataset may provide different manifestations of the same physiological event or medical condition, thus exhibiting a significant level of dependency. The dependencies between clinical variables should be explored, as they show great potential to improve the performance of diagnostic and prognostic prediction models [13], [14].

In addition, correlation (a form of probabilistic dependency) provides useful information to the prediction of missing values, improving the accuracy of the imputation process [6], [7], [15], [16]. Therefore, focus will be placed on the concept of correlation.

2.2.1 Correlation

Correlation is a measure of association between two variables, i.e. an indicator of how much a change in the magnitude of one variable is related to a change in the magnitude of another variable [17]. Although knowing the values of a variable A permits better

prediction of a correlated variable B, correlation does not necessarily assure causality [18]. A paradigmatic example that refutes this common misconception is the increase of both ice cream and fan sales in the summer, which results in a strong correlation between these two events. However, this association does not substantiate the hypothesis that eating ice cream causes people to purchase fans.

Hence, even though a causal relationship may occur between two variables, correlation analysis would not be sufficient to justify its existence. In the same way, it is possible to distinguish between regression and correlation: while the former expresses how the knowledge of A enables the prediction of B (one way association), the latter measures the strength of a mutual and symmetric association [19].

In the literature, correlation is most often used in reference to monotonic linear associations between variables, albeit non-linear relations can also be studied [17], [18]. Furthermore, correlation can also be described in terms of strength and direction. Assuming a monotonic relation between the correlated variables A and B, the correlation is positive (negative) if an increase in A is associated with an increase (decrease) in B.

Correlation coefficients are statistical measures of the degree of correlation between variables. There are a myriad of coefficients, each suitable for specific types of variables and with distinct underlying assumptions (e.g. linear association). Prior to presenting some of these coefficients, a taxonomy for the classification of variables will be discussed.

2.2.1.1 Types of Variable

There are various terminologies concerning the classification of variables, such as qualitative or quantitative, discrete or continuous, independent or dependent. The classification of interest for this work stems from the so-called scales of measurement: nominal, ordinal, interval, and ratio. This taxonomy and the properties of each scale permit the choice of the most appropriate statistical method for data analysis [20].

- **Nominal Variables** - Qualitative variables containing two or more labels (categories) without an intrinsic and undisputable order. Nationality, sex, blood group and cause of death are examples of nominal variables. Even if a nominal variable is numerically encoded, it makes no sense to perform mathematical operations on the numbers, as they have no meaning nor represent a specific ordering [21].

- **Ordinal Variables** - Similar to nominal variables, except for the possibility of establishing a natural ranking within the set of labels. For instance, opinions about a certain subject ("strongly agree" to "strongly disagree") exhibit a natural ordering and, therefore, are categories of an ordinal variable. The education level is another example. Mathematical operations on a numerically encoded ordinal variable are not meaningful as well, excluding comparisons between labels.

- **Interval Variables** - Quantitative variables measured on a scale without a "true zero point", i.e. where the measurement of 0 does not signify an absence of the attribute [20]. For example, temperature measured in degrees Celsius is an interval variable because 0 °C does not imply that there is no temperature.

- **Ratio Variables** - Quantitative variables that, in addition to all the properties of interval data, are measured on a scale with a "true zero point". Height, weight, heart rate, and temperature measured in Kelvin are examples of ratio variables.

Categorical data include two measurement scales: nominal and ordinal. Within the scope of ML, categorical data may require encoding procedures, such as one-hot encoding, which produces several binary variables (one per each category). Thus, it is important to include binary variables, i.e. categorical attributes consisting of two non-overlapping labels, in this discussion.

As for the interval and ratio scales, they measure numeric / metric data. These last designations are more prevalent in ML literature, since most, if not all techniques handle interval variables similarly to ratio ones.

2.2.1.2 Correlation Coefficients

Within this section, some of the most widely-used correlation coefficients will be presented and described. Table 2.1 was designed to provide a clearer overview of these correlation measurements across different variable types.

Table 2.1: Correlation coefficients across different variable types.

	Binary	**Nominal**	**Ordinal**	**Numeric**
Binary	Phi Coefficient, Cramér's V			
Nominal	Cramér's V	Cramér's V		
Ordinal	Rank Biserial	Rank Biserial Extension	Spearman, Kendall's Tau	
Numeric	Point Biserial	Point Biserial Extension	Spearman, Kendall's Tau	Pearson, Spearman, Kendall's Tau

1. **Pearson's Coefficient**

   The Pearson's coefficient, or Pearson's product-moment correlation coefficient, is one of the most used correlation measurements in medical research [22]. Commonly denoted by r, this coefficient measures the strength of a linear relationship between two numeric random variables X and Y, calculated by the following equation:

$$r = \frac{\text{cov}_{XY}}{\sigma_X \sigma_Y} \tag{2.1}$$

   where cov_{XY} is the covariance value between X and Y, and σ_X and σ_Y are the standard deviations of variables X and Y, respectively.

   The covariance assesses the joint variability of two random variables, but its value is dependent on the magnitudes of the variables. The Pearson's coefficient is a normalised covariance, scaled so that it ranges from -1 to +1, which respectively indicate a perfect negative and a perfect positive linear correlation [17]. A coefficient close to 0 refutes the existence of a linear relation between the variables, but never the existence of correlation, since the existing relationship may be non-linear. Figure 2.1 depicts constructed examples that illustrate how the Pearson's correlation coefficient reflects different relationships between variables.

   In order to draw proper conclusions regarding the degree of correlation, some assumptions have to be met. In fact, Pearson's correlation is only suitable for random numeric variables (interval and ratio) that follow a bivariate normal distribution [17], [18]. In addition to this constraint, this coefficient is fairly sensitive to outliers.

2. **Spearman's Rank Coefficient**

   The Spearman's rank correlation coefficient, usually denoted by r_s, is a rank-based Pearson's coefficient. Instead of using the actual values of two random variables to compute the coefficient, it uses their ranks as shown in the next equation:

$$r_s = 1 - 6 \times \frac{\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \tag{2.2}$$

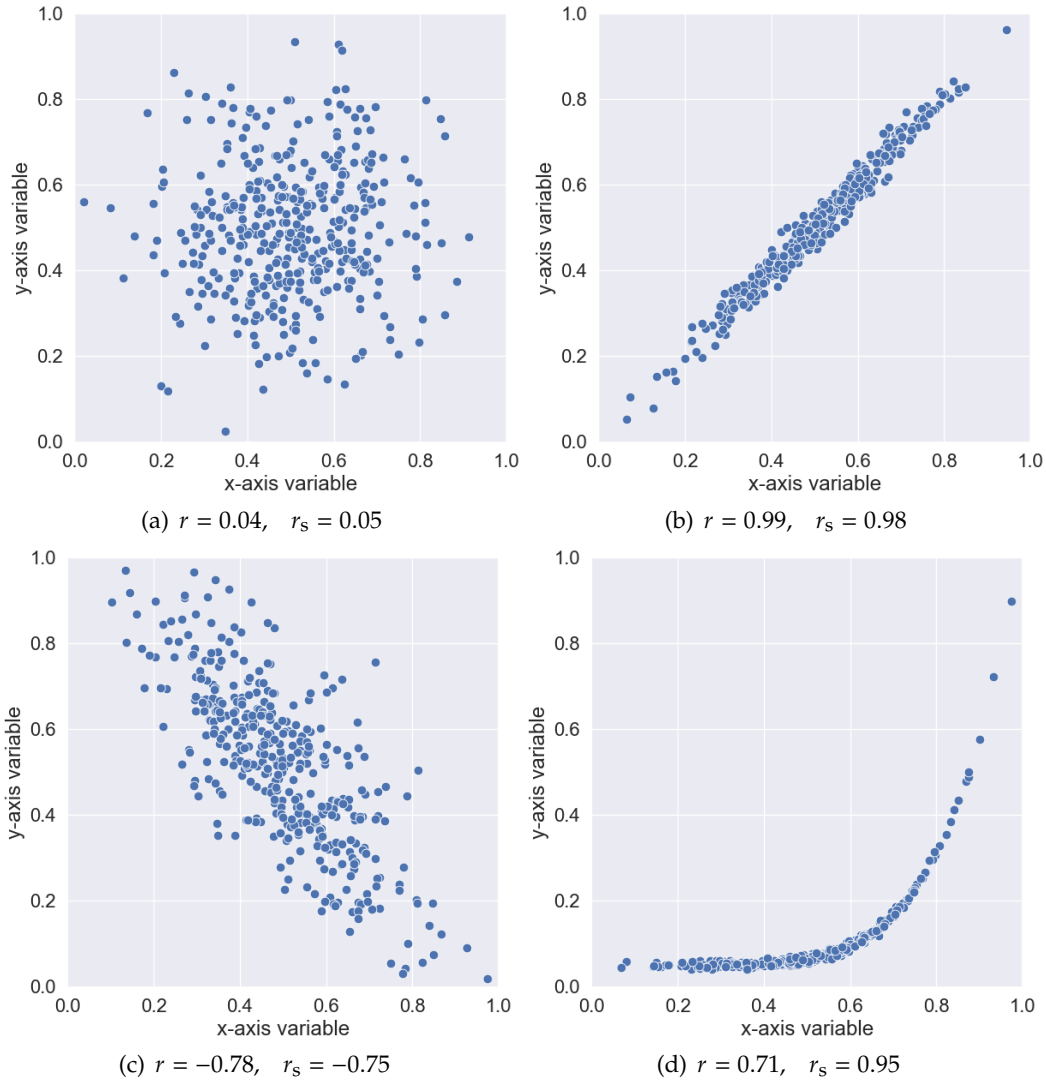
   where d_i is the rank difference of each pair of observations and N is the number of observations.

   Analogous to the Pearson's correlation, a Spearman's coefficient also ranges from –1 to +1. However, ranking of the data brings a few differences comparing to the Pearson's coefficient:

   - Spearman's correlation is suitable for all except nominal variables, as both numeric and ordinal can be ranked. Furthermore, it can be used even when two numeric variables do not follow a bivariate normal distribution [18];

- This coefficient measures the strength of a strictly monotonic relationship between two variables, which does not need to be linear due to the use of ranks. A coefficient close to 0 indicates the absence of a monotonic relationship, while the extreme values of -1 and +1 describe a perfect negative and perfect positive monotonic association, respectively;
- The robustness to outliers is greater [17].

Figure 2.1 allows a better understanding of when the values obtained through the Pearson's and Spearman's correlation coefficients are distinct, as well as the situations in which one is more appropriate than the other.



(a) $r = 0.04$, $r_s = 0.05$ (b) $r = 0.99$, $r_s = 0.98$

(c) $r = -0.78$, $r_s = -0.75$ (d) $r = 0.71$, $r_s = 0.95$

Figure 2.1: Pearson's and Spearman's correlation coefficients, r and r_s, for different relationships between two normally-distributed variables X and Y: (a) X and Y are practically independent, (b) X and Y have a strong positive linear relationship, (c) X and Y have a moderate negative linear relationship, (d) X and Y have a strong positive exponential relationship. Based on [17].

3. **Kendall's Tau Coefficient**

   The Kendall's Tau correlation coefficient, denoted by τ, is another rank correlation measurement, considered an extension of the Spearman's coefficient [23]. Hence, it quantifies the strength of a monotonic relationship between two random variables, X and Y, which can be ordinal or numeric.

   Contrarily to the Spearman's correlation, Kendall's Tau does not evaluate the association between two rankings. Instead, it is a measure of how many transpositions are needed to get both variables in the same order [18]. In fact, its calculation is based on the number of concordant and discordant pairs of observations. A pair of observations (x_i, x_j) and (y_i, y_j), with $i < j$, is said to be concordant if the unit ranking higher on X ranks higher on Y, i.e. $x_i < x_j$ and $y_i < y_j$ or $x_i > x_j$ and $y_i > y_j$ [24]. A pair is discordant otherwise.

   This coefficient has some variations and the most widely-used is Kendall's Tau-b, denoted by τ_b, which accounts for pairs with tied ranks. When ranking a variable, a tied rank is a rank shared by two or more units that have the exact same value on that variable [25]. Kendall's Tau-b is calculated through the following equation:

$$\tau_b = \frac{n_c - n_d}{\sqrt{[n(n-1)/2 - T_X] \times [n(n-1)/2 - T_Y]}} \tag{2.3}$$

   where n is the number of observations, n_c the number of concordant pairs, n_d the number of discordant pairs, and T_X and T_Y are the number of tied pairs on X and Y, respectively.

   Similarly to the previous coefficients, Kendall's Tau-b also ranges from -1 to +1, which respectively indicate a perfect negative and a perfect positive association between the two rankings.

   In comparison to Spearman's, this coefficient is more suitable for small datasets, although the former is still more widely used [23]. Furthermore, Kendall's Tau frequently yields smaller correlation values than Spearman's rank coefficient [26]. Lastly, Puth et al. [27] stated that, in the presence of any tied ranks, Spearman's correlation should be considered superior to Kendall's Tau.

4. **Point Biserial Coefficient**

   The point biserial correlation coefficient, represented by r_{pbi}, measures the strength of association between a binary nominal variable Y and a numeric variable X [28]. The following equation can be used to compute this coefficient:

$$r_{pbi} = \frac{\bar{X}_1 - \bar{X}_0}{\sigma_X} \sqrt{p_1 p_0} \tag{2.4}$$

   where p_1 and $p_0 = 1 - p_1$ are respectively the proportion of units with $Y = 1$ and $Y = 0$, $\bar{X}_1$ and $\bar{X}_0$ are respectively the means of X given $Y = 1$ and $Y = 0$, and σ_X is the standard deviation of X.

This measurement ranges from -1 to +1 and is based on the Pearson's coefficient [29]. Thus, its interpretation is similar: a positive (negative) value indicates that the two variables are related positively (negatively), and a higher coefficient represents a greater degree of association.

There is a generalization of this coefficient suitable for when Y is a nominal variable with more than two labels. The point biserial extension coefficient assumes that Y follows a multinomial distribution and that the conditional distribution of X for fixed Y is multivariate normal [30].

5. **Rank Biserial Coefficient**

   The rank biserial correlation is similar to the point biserial correlation, except that the numeric variable X is ranked, i.e. it can be ordinal [29]. This coefficient also has an extension for when Y is a nominal variable with more than two labels.

6. **Phi Coefficient**

   The Phi coefficient, denoted by ϕ, is the equivalent to Pearson's coefficient that measures the linear correlation between two binary variables X and Y. It can be calculated through the Pearson's chi-square goodness-of-fit statistic, or simply chi-square statistic, for the 2×2 contingency table of the two variables:

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{2.5}$$

   where χ^2 is the chi-square statistic and N is the number of observations.

   While the chi-square statistic expresses the statistical significance of the relationship, ϕ describes the strength of the association [19]. The 2×2 contingency table is helpful to determine the sign of the coefficient: the table for a positive correlation should be closer to a diagonal matrix, with most of the data falling along the diagonal cells; the opposite happens for a negative correlation. The sign is the same as $n_{11}n_{00} - n_{10}n_{01}$ where n_{ij} are the number of observations where $X = i$ and $Y = j$, with $i, j = \{0, 1\}$.

   Similar to the Pearson's correlation, the Phi coefficient also ranges from -1 to +1, which respectively represent a negative and a positive perfect association. A value close to 0 indicates the absence of a linear association.

7. **Cramér's V**

   Cramér's V is an extension of the Phi coefficient for larger contingency tables, as it measures the strength of linear association between two nominal variables with two or more labels [29]. Its formula is also based on the chi-square statistic, as shown below:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}, \qquad k = \min(r, c) \tag{2.6}$$

where χ^2 is the chi-square statistic, N is the number of observations, r and c are respectively the number of rows and columns of the contingency table, and k is the lesser number of categories of either variable. When $k = 2$, Equation 2.6 becomes equivalent to Equation 2.5, which concerns the calculation of the Phi coefficient.

Cramér's V ranges from 0 to 1, with 0 indicating the absence of an association between the variables and 1 representing a perfect correlation.

Bergsma [31] observed that for small sample sizes, the empirical value of V potentially overestimates the strength of association. To overcome this limitation, a bias correction was presented:

$$\tilde{V} = \sqrt{\frac{\tilde{\phi}^2}{\tilde{k} - 1}}, \qquad \tilde{k} = \min(\tilde{r}, \tilde{c}) \tag{2.7}$$

where

$$\tilde{r} = r - \tfrac{1}{N-1}(r-1)^2,$$
$$\tilde{c} = r - \tfrac{1}{N-1}(c-1)^2,$$
$$\tilde{\phi}^2 = \max\left(0, \phi^2 - \tfrac{1}{N-1}(r-1)(c-1)\right).$$

2.3 Machine Learning

While AI comprehends all programs and algorithms that learn like humans, ML is the branch of AI that allows systems to automatically learn from past data. ML algorithms make decisions or predictions without being explicitly programmed and improve their performance with experience [32], [33]. This ability has led to a widespread use of ML in areas such as health care, cybersecurity, agriculture, and many others.

2.3.1 Machine Learning Taxonomy

ML models can be divided into four categories according to the learning process:

- **Supervised Learning (SL)** - A learning process is supervised if the instances (examples) used for training the model are labelled, i.e. it is provided information regarding their target values or categories. Based on the patterns and relationships between these instances and their respective labels, the system estimates a mapping function and uses it to make predictions about unseen input data. SL models are generally used to solve two types of problems: regression and classification [34]. In the former, the target variable is a continuous value, whereas in the latter it is a discrete value or category. Linear and Ridge regressions are some examples of regression models. A classification problem may be binary, when there are only two possible outcomes, or multiclass, when it deals with more than two classes. The most known classification methods include Decision Trees (DTs), Random Forests

(RFs), K-Nearest Neighbours (KNNs), Naive Bayes (NB) Classifiers and Support Vector Machine (SVM) Classifiers [35]. SL models can also be segmented into discriminative and generative, depending on how the algorithms reach their prediction. A discriminative model focuses on the decision boundary between the different classes. Examples of these models are: Logistic Regressions, KNN and SVM. A generative model explicitly models the detailed characteristics of each class, such as its probabilistic distribution. The Bayes Theorem is used as the basis for designing these learning algorithms, which include models such as the NB.

- **Unsupervised Learning** - In contrast to SL, unsupervised models learn through unlabelled data. These systems discover similarities and differences in the input instances, identifying hidden patterns without the need for human interference [33]. The most common tasks include clustering, dimensionality reduction, association rule discovery and anomaly detection.

- **Semi-supervised Learning** - The learning process is said to be semi-supervised when it uses both labelled and unlabelled data, and can be regarded as a hybridisation of supervised and unsupervised learning. Semi-supervised models aim to provide a better prediction than that obtained through labelled or unlabelled data alone [33].

- **Reinforcement Learning** - In a reinforcement learning problem, an autonomous agent learns the optimal course of action, i.e. behaviour, through trial-and error interactions with a dynamic environment [36]. The learning process is driven by a numerical reward, and the agent must attempt different sequences of actions and progressively favour those which maximise this reward, as it reflects in a better performance.

2.3.2 Machine Learning Pipeline

The sequence of processes involved in the development of a ML model is known as ML pipeline. There are three key stages of development: pre-processing, training, and performance evaluation. A schematic representation of a general ML pipeline is show in Figure 2.2. A description of each of these phases will be presented in the following subsections, as well as some additional steps. These topics will be addressed from a SL perspective and greater focus will be placed on the concepts used in this work.

2.3.2.1 Pre-processing

The majority of real-world datasets are often incomplete, inconsistent, and contaminated with noise due to their heterogeneous origin [4]. These issues have a deleterious effect on the learning process, and therefore it is essential to ensure data quality, i.e. ensure a clean, well organised and normalised dataset. The pertinence of a pre-processing stage lies precisely in this necessity. Within this scope, practices such as data cleaning, feature extraction, feature selection, categorical encoding, and data scaling will be addressed.

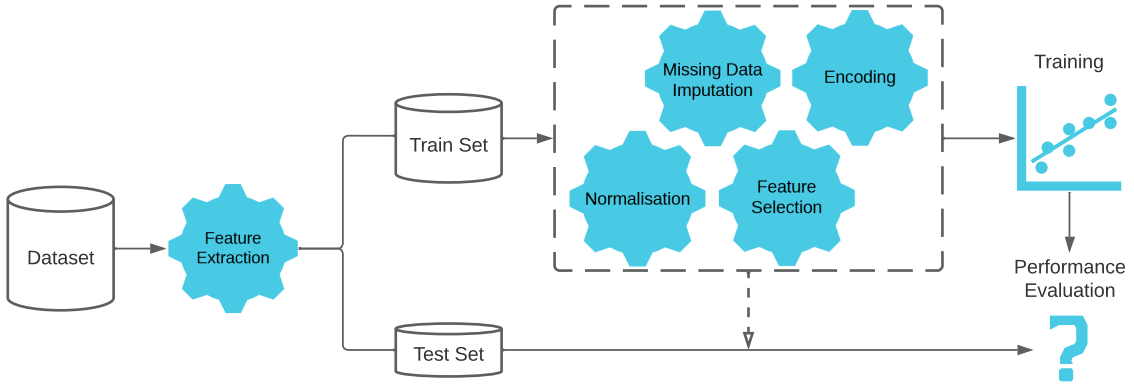


Figure 2.2: Development of a ML model.

Data cleaning is the detection and handling of corrupt or inaccurate data. This process may include missing data imputation, outlier detection, and noise removal [37]. A wide variety of data analysis methods becomes inappropriate or difficult to apply on incomplete datasets, thus revealing the need to deal with missing data [38]. This issue will be discussed in more detail in Section 2.4, where additional problems associated with missing values are presented, as well as various approaches to tackle this challenge.

The feature extraction process transforms data into a meaningful numerical representation, i.e. a representation that captures and highlights the patterns in the data and its structure [39]. On the other hand, feature selection consists of detecting the most relevant features and discarding irrelevant and redundant data, i.e. features that neither provide additional nor useful information [40]. Hence, this procedure selects which subset of features will be used to train the ML model. The implementation of these two practices typically leads to an enhanced performance of predictive models.

Although categorical data is present in a significant amount of real-world applications, the grand majority of ML models are developed for numeric variables [41]. Therefore, each categorical feature needs to be encoded, i.e. transformed into their numerical counterpart. Ordinal encoding and one-hot encoding are amongst the most widely-used techniques of categorical encoding:

- **Ordinal encoding** - Each category is represented by a different integer, which typically ranges from 0 to $n-1$, where n is the number of categories. Although this procedure does not change the cardinality of the variable, and thus does not add new columns to the dataset, it introduces an ordering that is absent in nominal variables. This encoding might be suitable for representing growing levels of severity of a medical condition, for example.

- **One-hot encoding** - A variable, containing n categories, is transformed into n binary columns, each encoding a different category. In the resulting binary columns, a value of 1 indicates that the sample in question belongs to the encoded category, and the value of 0 indicates otherwise. Even though this procedure overcomes the previous limitation for nominal variables, it can considerably increase the dataset's dimension and originate a sparse set of features.

Data scaling is the process of remapping the data variables from their original format to an unified format for all features, adjusting for range and / or distribution [39]. This transformation prevents bias in the results produced by a ML model [4]. For instance, in a classifier based on the Euclidean distance, features with a larger range of values would be more determinant (have a larger contribution) if the dataset was not previously scaled. There are two common approaches for data scaling:

- **Normalisation** - Converts each value x of a certain feature X to a value x_{norm} in the range $[low, high]$:

$$x_{\text{norm}} = low + \frac{(high - low)(x - X_{\text{min}})}{X_{\text{max}} - X_{\text{min}}} \tag{2.8}$$

  where X_{min} and X_{max} are respectively the minimum and maximum values of X. The variables low and $high$ are usually either -1 and 1, or 0 and 1.

- **Z-score normalisation/ Standardisation** - Forces each feature to have the properties of a standard normal distribution, with a mean of 0 and a unit standard deviation. Each value x is replaced with its standard score x_{stand}, given by the next equation:

$$x_{\text{stand}} = \frac{x - \mu}{\sigma} \tag{2.9}$$

  where μ and σ are the statistical mean and standard deviation, respectively.

2.3.2.2 Training

In any ML pipeline, it is fundamental to split the dataset into train and test subsets, each serving a different purpose. To avoid data leakage, procedures such as missing data handling, feature selection, categorical encoding and data scaling are first performed on the train set. The knowledge obtained through that subset is used to transform the test set. Then, the ML model learns upon information from the train set and estimates a mapping function, based on the common patterns that denote the samples of the same class [34].

The main objective of this work does not directly imply the development and optimisation of the ML models usually involved in this step, since the focus is on the data itself, more specifically on the precision of the imputation procedure. A great part of the success of a ML model is due to the quality of the data it is trained upon. Therefore, by evaluating the performance of models trained upon imputed data, we can indirectly evaluate the missing data handling approach used for imputation. Moreover, it is important

to understand how ML algorithms work, in order to perceive how they can be affected by the imputation of missing values. For these reasons, a succinct description of the RF, SVM and NB algorithms will be presented, as they are some of the most commonly used models.

1. **Random Forest**

   A RF is based on DT classifiers, and thus this algorithm will also be briefly described. A DT is composed of nodes, or subdivisions, that guide the sorting of a given instance from the DT's root to a leaf, i.e. a node that provides the classification. Every non-terminal node contains a split based on a single feature, and each branch descending from that node corresponds to one of the possible values (or interval of values) for this attribute [42]. The non-terminal nodes should apply the most beneficial division from a learning perspective. Hence, each division must meet a "goodness of split" criterion, and the most widely-used are the Gini Impurity Index and Information Gain [43].

   Returning to the RF algorithm, it is an ensemble technique that fits several simple DTs in parallel and independently, using different subsets of training samples. The multiple outputs are combined to reach a single result, thus increasing prediction accuracy [33].

2. **Support Vector Machine**

   The aim of a SVM algorithm is to find the optimal hyperplane in a high-dimensional feature space, i.e. the hyperplane that best separates the data points [44]. The minimum distance between a hyperplane and the data points nearest to it is called the margin, and it is widely used to measure the efficiency of the separation enforced by a hyperplane. As depicted in Figure 2.3, the optimal hyperplane achieves the maximum margin of separation between the two classes. Although it was first proposed as a two-class discriminant function, extensions of this method have been developed, which include nonlinear SVMs.

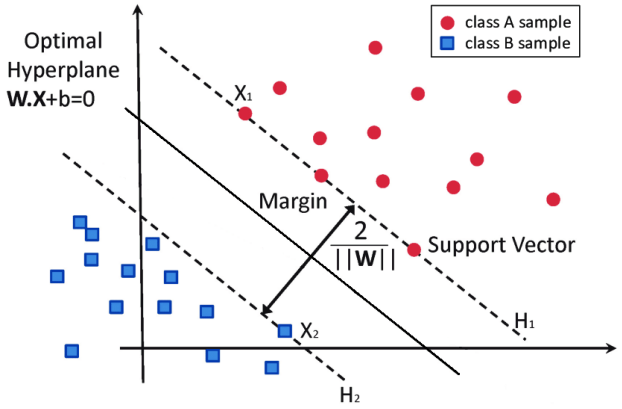


Figure 2.3: Binary classification performed by a SVM algorithm. Adapted from [45].

3. **Naive Bayes**

   A NB classifier is a probabilistic model whose design stems from the Bayes Theorem. Therefore, this ML algorithm assumes that the features are independent within a given class, and calculates the final classification for a certain instance through the following equation:

$$\hat{y} = \underset{j \in \{1,\dots,m\}}{\operatorname{argmax}} \prod_{k=1}^{f} P(x_k|y_j)P(y_j) \tag{2.10}$$

   where $\hat{y}$ is the final prediction for sample x, y_j is the jth class out of the m possible classes, x_k is the kth feature of sample x within a total of f attributes, $P(x_k|y_j)$ is the probability of x_k given a class y_j, and $P(y_j)$ is the probability of class y_j.

When building any ML model, varying the values of the hyperparameters, i.e. parameters that define the model architecture, may lead to distinct performances. Examples of hyperparameters of the RF classifier are the number of trees and the "goodness of fit" criterion. Hyperparameter tuning is often performed, as it allows the identification of the set of values that results in an optimal model architecture [46]. The main three techniques used to perform hyperparameter optimisation are the following:

- **Grid Search** - This approach performs an exhaustive search of the optimal configuration of hyperparameters, whose values belong to a given fixed set. In other words, a list of values is specified for each chosen parameter, the model is trained upon each possible combination, and the one that produced the best performance results is selected as optimal.

- **Random Search** - In this strategy, instead of specifying a list of discrete values for every hyperparameter, a statistical distribution must be defined for each one. Then, combinations of values are randomly selected to train the model, and the one that produced the best performance results is selected as optimal. Contrarily to grid search, not all possible configurations are tested.

- **Bayesian Optimisation** - This approach does not treat each configuration of hyperparameters independently, as the last two did. In fact, each combination of values is determined considering previous results, which avoids unnecessary iterations.

After the best algorithm and set of hyperparameters is defined, the model is apt to be trained upon the whole training subset.

2.3.2.3 Performance Evaluation

After this training process, the test set is given as an input to the model in order to evaluate its performance on unknown data. If the results are appropriate, the model is considered ready to be implemented and evaluated in a real-world scenario. If not, the development

process must be adjusted or even reformulated, as the training process must be repeated in different conditions to achieve better results.

The performance evaluation is dependent on the data-splitting process, and several techniques can be adopted considering each specific context:

- **Holdout set** - The dataset is partitioned into two mutually exclusive subsets, the training and the test set. The former is larger and is given as an input to the model, while the latter is used for evaluation.
- **k-fold Cross Validation** - The dataset is divided into k mutually exclusive folds of approximately equal size. The model is trained k times: in each iteration, one of the folds serves as the test set and the remaining are used as the training set. A particular case of this strategy is the leave-one-out method, where one sample is left out for evaluation at each iteration.
- **Bootstraping** - In this technique, the training set is created by sampling with replacement n instances from the dataset of size n. The remaining samples constitute the test set and are used for evaluation.

The evaluation metrics are used to assess the generalisation ability of a trained model, i.e. its performance when tested against unseen data [47]. Moreover, these metrics can be used for model selection as they enable a simple comparison of different classifiers' performance.

In binary classification there is always a positive and a negative class, which can represent the binary output values (1, 0) or (true, false), respectively [35]. In multiclass classification, the positive class is one chosen from all the possibilities and the negative class is an agglomeration of the remaining. This division into positive and negative classes allows the definition of the following key concepts:

- True Positive (TP) – Number of instances correctly classified as positive.
- False Positive (FP) – Number of instances incorrectly classified as positive.
- True Negative (TN) – Number of instances correctly classified as negative.
- False Negative (FN) – Number of instances incorrectly classified as negative.

The definition of some of the most commonly used evaluation metrics relies on the aforementioned concepts.

1. **Confusion Matrix**

   In a classification problem, a confusion matrix is the standard structure to represent the model's performance. For problems with N target classes, the confusion matrix will be a $N \times N$ matrix. This representation summarises the number of correct

and incorrect predictions for each target class, showing the distribution of the misclassified test samples. Figure 2.4 shows the confusion matrix for a binary classification problem.

		Actual Classes	
		Positive (1)	Negative (0)
Predicted Classes	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 2.4: Confusion Matrix of a binary classifier.

A perfect classifier (binary or multiclass) has a diagonal confusion matrix, as it only makes correct predictions.

2. **Accuracy**

   The accuracy metric consists of the fraction of correctly predicted classes over the total number of instances within the test set. In a binary classification problem:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.11}$$

   In multiclass classification, the accuracy is obtained by dividing the sum of elements in the main diagonal of the confusion matrix with the total number of instances.

3. **Recall**

   The recall, also known as sensitivity, measures the percentage of real positive instances that are actually classified as positive, i.e. the True Positive Rate (TPR). Its formula for a binary classification problem is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{2.12}$$

   For multiclass classification, the recall for each class is calculated through Equation 2.12, and then an arithmetic mean of these values is used to obtain the macro average recall [48]. There is also a micro average score, but the macro average is more suitable when all classes are equally important (even in imbalanced data).

4. **Precision**

   The precision metric measures the ratio of predicted positive instances that are in fact real positive samples. In binary classification:

$$Precision = \frac{TP}{TP + FP} \tag{2.13}$$

As for a multiclass classification problem, the macro average precision is calculated in a similar fashion as the previous metric, i.e. through an arithmetic mean of the precision values for each class.

5. **F1-Score**

   The F1-Score represents the harmonic mean between the precision and recall metrics [47]. For binary classification problems:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.14}$$

   And for multiclass classification, the macro F1-Score can be calculated through the following equation [48]:

$$Macro\ F1 - Score = \frac{2 \times MacroAveragePrecision \times MacroAverageRecall}{MacroAveragePrecision + MacroAverageRecall} \tag{2.15}$$

   The macro-averaged F1-Score was preferred for this dissertation, as it considers all classes to be equally important, even on imbalanced data. A weighted average would give greater importance to classes with more instances, which is the opposite of what is intended.

6. **Specificity**

   The specificity measures the proportion of real negative instances that are actually classified as negative, i.e. the True Negative Rate (TNR). In binary classification:

$$Specificity = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.16}$$

   In a multiclass classification problem, the macro average specificity is the arithmetic mean of the specificity values for each class.

7. **Area Under the Receiver Operating Characteristic (AUROC)**

   The output of a ML classifier may be given as a numeric value that represents the degree to which an instance is a member of a class, i.e. a probabilistic score ranging from 0 to 1 [49]. After defining a cut-off value, it is possible to generate a binary outcome: when the probabilistic score is above the threshold, the output is a positive case; otherwise, it is a negative case. The Receiver Operating Characteristic (ROC) curve is a two-dimensional graph in which the sensitivity, i.e. the TPR, is plotted against $1 -$ specificity, i.e. the False Positive Rate (FPR), for different cut-off values. This curve describes the evolution of correctly predicted positive cases as the number of false positives increases. Figure 2.5 depicts the ROC curves for a perfect, a generic, and a random classifier.

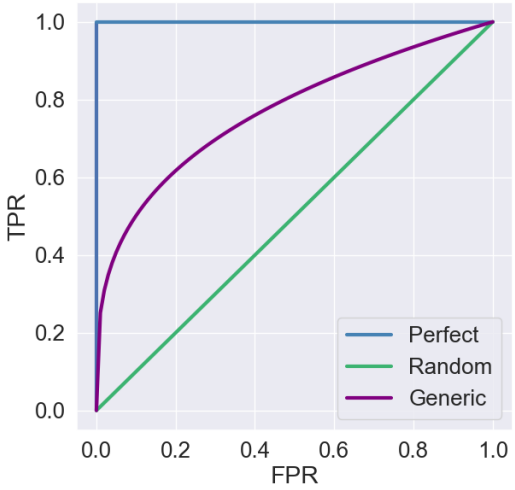


Figure 2.5: ROC curves for a perfect, a random, and a generic classifier.

A perfect classifier would have a TPR of 1 for any non-zero values of FPR, and thus the ROC curve runs along the vertical y-axis to the point (0,1), and then goes horizontally until it reaches the point (1,1) [50]. The ROC curve of a random classifier, which is equally likely to produce a false positive or a true positive, is a diagonal line going from the origin to the point (1,1). A generic classifier usually exhibits a ROC curve that lies between the previous two.

The AUROC is a performance measurement for classification problems, with a higher value indicating that the model is closer to an ideal situation in which all positive instances are correctly classified. This ability is of particular interest within the scope of clinical diagnosis [51].

In multiclass classification, the ROC curve can be plotted for all classes individually, with the remaining ones serving as negative cases.

2.4 Missing Data

Little and Rubin [52] defined missing data as "unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value". This absence of information poses a recurring predicament to several research fields, including medical sciences.

There are various factors that could cause missing data, such as faulty equipment, incorrect measurements, drop-out in studies and human errors [53], [54]. In a clinical context, Wells et al. [55] identified two main causes for missing information: lack of collection, e.g. when a patient's blood pressure is not measured, and lack of documentation, e.g. when their blood pressure is measured but not registered in the medical record.

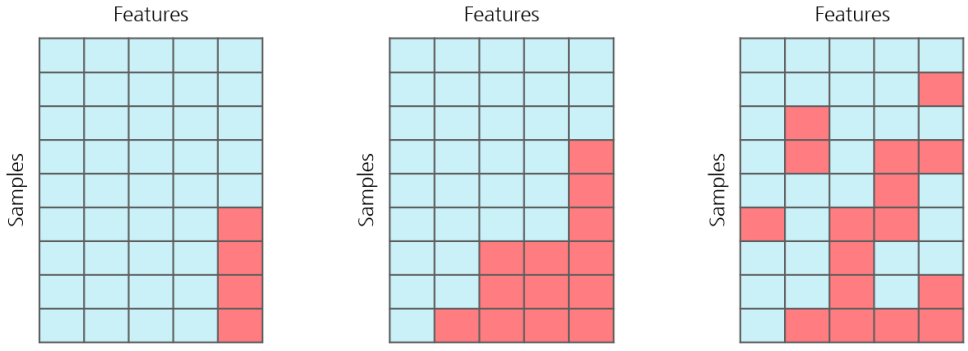
Missing values bring numerous consequences for a research work. Firstly, it is expected a decrease in the information extracted from the data, which may result in a loss of efficiency and precision, and lead to conclusions statistically less strong [53], [56]. Furthermore,

common data analysis methods often become inappropriate or difficult to apply on incomplete datasets [38]. The gravity of these problems is aggravated by an increase in the MR, i.e. the percentage of missing data [6].

Little and Rubin [52] argued that some unobserved values cannot be regarded as missing data, and thus should be handled differently. For example, consider a study with two phases, conducted one year apart, that evaluates the well-being of hospitalised patients. It is likely that the second phase will have more missing information, as the patients may no longer be alive. However, well-being is not a meaningful concept for people who are not alive, and therefore the aforementioned definition of missing data does not apply.

A nearly universal classification system for missing data problems was established by Rubin [57], who pioneered the study on how the process that causes missingness affects the analysis of data. Within this scope, Little and Rubin [52] distinguished missingness pattern from missingness mechanism, terms which are often used interchangeably. The former refers to the configuration of observed and missing values in a data matrix, whereas the latter describes possible relationships between missingness and measured values.

Figure 2.6 shows the three missing data patterns, presented by Little and Rubin [52], that appear most in the literature. In an univariate pattern, as opposed to a multivariate one, the missingness is restricted to a single variable. If the variables in the data can be arranged so that the configuration of observed and missing values resembles a staircase, then the pattern is monotone. When this setup is not possible, the missingness pattern is classified as general.



(a) Univariate (b) Monotone (c) General

Figure 2.6: Examples of missing data patterns: (a) Univariate, (b) Monotone, (c) General. Observed and missing values are represented respectively by the colours blue and red.

2.4.1 Missingness Mechanisms

Little and Rubin [52] categorised the missingness mechanisms as Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing not at Random (MNAR). Although these mechanisms do not offer a causal explanation for the missingness, they describe generic relations between data and missingness, which are important to understand

when choosing the appropriate method to handle the missing values [58], [59].

Before delving into each mechanism, it is important to present some concepts used for their definition. Following Little and Rubin [52], let $Y = (y_{ij})$ denote an $(n \times K)$ rectangular dataset without missing values, i.e. a complete dataset, where y_{ij} is the value of variable Y_j for unit or observation i. If Y contains missing data then the missingness indicator matrix $M = (m_{ij})$ is defined, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed.

The formal description of the missingness mechanisms relies on the conditional distribution of m_i given y_i, hereby represented as $f_{M|Y}(m_i|y_i, \phi)$, where ϕ denotes unknown parameters.

2.4.1.1 Missing Completely at Random

In a MCAR mechanism the missingness is completely unrelated to the values of the data, missing or observed. This mechanism verifies the equality

$$f_{M|Y}(m_i|y_i, \phi) = f_{M|Y}(m_i|y_i^*, \phi) \tag{2.17}$$

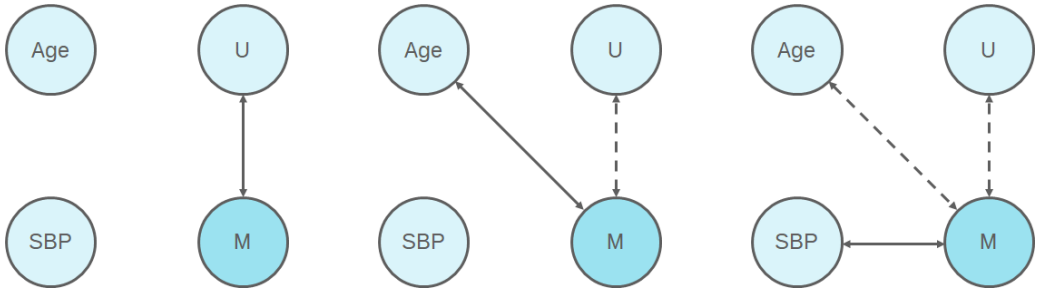
for all i and any distinct values (y_i, y_i^*) in the sample space of Y, as formally defined by Little and Rubin [52]. Accordingly, Equation 2.17 acknowledges that the probability of missingness on a variable Y_j is not dependent on other measured variables nor on the values of Y_j itself [58].

Table 2.2 was designed to better illustrate these concepts. Consider a routine medical checkup, where the provider registers the patient's age and collects standard measurements such as weight, height, blood pressure, and others. The variables Age and Systolic Blood Pressure (SBP) are reported in Table 2.2, as well as the three missingness mechanisms applied to the latter. Figure 2.7 depicts a schematic representation of this example.

Table 2.2: Patient's age and SBP values, measured during a routine medical checkup. The three missingness mechanisms were applied to the variable SBP.

Age	**SBP (mmHg)**			
(years)	**Complete**	**MCAR**	**MAR**	**MNAR**
11	110	110	?	?
16	126	?	?	126
21	108	108	?	?
29	133	133	?	133
35	96	96	96	?
40	100	?	100	?
43	152	?	152	152
56	116	116	116	116
67	148	148	148	148
71	140	?	140	140

If the provider forgets to record the SBP in an haphazard manner, i.e. with no relation to the patient's age nor to the value of the reading itself, then the variable will be MCAR. This mechanism is represented on the third column of Table 2.2, which shows that the missing values are not confined to a particular location in the Age or SBP distributions. Additionally, note that the MCAR schematic in Figure 2.7(a) only exhibits an association between the missingness and unmeasured variables, since there is no relation between the missingness and the data within this mechanism.



(a) MCAR (b) MAR (c) MNAR

Figure 2.7: Schematic representation of the missingness mechanisms in the given example of a routine medical checkup: (a) MCAR, (b) MAR, (c) MNAR. The letters M and U represent, respectively, the missingness indicator matrix and a set of unmeasured variables. The solid arrows indicate statistical associations, while the dashed ones denote possible relationships. Based on [58].

Another example of a MCAR mechanism happens when a patient's laboratory test results are missing because the samples were incorrectly processed. Also, if a subject drops out of a longitudinal study for reasons unrelated to any factors under study, the resulting missing values are considered MCAR [60].

2.4.1.2 Missing at Random

In a MAR mechanism the missingness is related to the observed values of the data, but not to the missing ones. As per the nomenclature adopted by Little and Rubin [52], let $y_{(0)i}$ and $y_{(1)i}$ denote the observed and missing elements of y_i, respectively. This mechanism verifies the equality

$$f_{M|Y}(m_i|y_{(0)i}, y_{(1)i}, \phi) = f_{M|Y}(m_i|y_{(0)i}, y^*_{(1)i}, \phi) \tag{2.18}$$

for all i and any distinct values $(y_{(1)i}, y^*_{(1)i})$ in the sample space of $Y_{(1)}$ [52]. In conformity with Equation 2.18, the probability of missingness on a variable Y_j depends solely on the values of other measured variable (or variables), but not on the values of Y_j itself [58].

Returning to the example illustrated in Table 2.2. If the provider deliberately does not measure the SBP of patients younger than 30 years old, then there will be a relation between the missingness and the variable Age. Consequently, the missing values are restricted to the lower segment of the Age's distribution, as represented on the fourth

column of Table 2.2. Since there is no dependency between the missingness and the values of the variable SBP, this is a case of a MAR mechanism.

The schematic in Figure 2.7(b) depicts an association between the missingness and the observed values, as well as a possible linkage between the missingness and unmeasured variables. However, the latter is not relevant to Rubin's theory, which exclusively studies the relationships between missingness and measured data, and thus will not be included in the current analysis.

The MAR mechanism may also appear in, for example, a survey examining depression. Suppose the male subjects are less likely to answer questions about depression severity than the female ones. As the missingness is dependent on a measured variable, i.e. the subject's sex, but not on the severity of their depression, then the data is MAR [61].

2.4.1.3 Missing not at Random

Finally, in a MNAR mechanism the missingness is related to the unobserved data. According to Little and Rubin [52], the distribution of m_i depends on the missing elements of y_i, i.e. Equation 2.18 does not apply for some unit i and some values $(y_{(1)i}, y^*_{(1)i})$. This is the only mechanism that permits an association between the probability of missingness on a variable Y_j and the values of Y_j itself. Additionally, the missingness can depend on the observed values of the data, as long as it still relates to the missing ones [58].

Consider, once again, the example represented in Table 2.2. Suppose the provider deliberately does not register SBP values under 115 mmHg. As shown on the last column of Table 2.2, the missing values are limited to the lower segment of the SBP's distribution. This indicates a relation between the missingness and the unobserved data, which is consistent with a MNAR mechanism.

Accordingly, the MNAR schematic in Figure 2.7(c) shows an association between the missingness and the unobserved values, along with a possible connection between the missingness and unmeasured variables. This last connection will not be included in the current analysis, for the aforementioned motives. Although in this particular example there is no evident association between the observed data and the missingness, nothing precludes this association from also being present in a MNAR mechanism.

Consider, as a final example, a questionnaire examining lifestyle and health habits. Subjects with a substance use disorder are more likely to refuse to answer questions about drug usage. Even though drug usage may depend on measured and observed variables, such as age or monthly income, the missing values depend on the usage itself and, therefore, the values are MNAR.

2.4.1.4 Assessing Missingness Mechanisms

Several methods have been developed to test if an attribute is under the MCAR mechanism [62]–[64], but none of these methods is able to provide definite evidence [58], [65]. For example, suppose a certain variable V is MCAR, and that the remaining variables of the

dataset are split into two subgroups: one where V is observed and another where it is missing. A simple test consists of performing group mean comparisons, because it is expected that these two subgroups share the same mean vector (recall that the observed values in V are a random sample of an hypothetically complete dataset). However, the remaining missingness mechanisms can also generate subgroups with equal means, which shows that this is not a conclusive test.

The practical problem with MAR and MNAR mechanisms lies in not knowing the real values of the missing entries. As a consequence, there is no way to confirm if the missingness in a certain variable is related to the unobserved data (MNAR) or is solely a function of other measured variables (MAR) [66].

On a final note, these missingness mechanisms should not be regarded as a predetermined trait of the dataset, as they can be affected by the performed analysis [58], [67]. Ponder the previous MAR example, concerning a survey that evaluates depression. If the variable Sex was left off the analysis model, then this would no longer be a MAR case. Instead, since the cause of missingness is now an unmeasured variable, this would be a MCAR analysis.

2.4.2 Missing Value Handling Approaches

When addressing the challenges imposed by missing data, researchers have developed a myriad of methods to overcome this issue. These techniques can be divided into three categories, according to the adopted strategy [52], [53]: deletion, imputation and model-based methods.

2.4.2.1 Deletion Methods

The deletion methods, or missing data ignoring techniques, are procedures based on completely recorded units. The deletion process can be executed in two ways: listwise deletion and pairwise deletion. The former discards the units with one or more missing values, performing a so-called complete-case analysis. The latter, on the other hand, omits units on an analysis-by-analysis basis, with its prototypical application being the use of different complete subsets to determine each element in a correlation matrix [58].

Deletion is undoubtedly the simplest strategy and thus widely employed in many areas of scientific research [68]. Despite its convenience, discarding incomplete units induces a potential loss of information, aggravated by an increase in the MR or in the number of variables. Moreover, this procedure may introduce bias in the analysis, particularly when the mechanism is not MCAR, as the remaining units are unrepresentative of the hypothetically complete dataset [52], [58]. Since the benefits do not counterbalance all the limitations, this strategy is inadvisable in most cases, except when the proportion of missing data is low.

2.4.2.2 Imputation Methods

The key purpose of an imputation technique is to fill in, i.e. replace, the missing elements by some predicted values, usually estimated from the observed data. In an univariate imputation approach, only the observed values of the missing variable are considered for the prediction. Contrarily, multivariate imputation presupposes including other variables in the estimation of the missing value. Additionally, imputation methods can be grouped into two categories: single imputation and Multiple Imputation (MI).

In single imputation, a sole replacement value is computed for each missing data point. The most frequent single imputation techniques in the literature include mean imputation, regression imputation, and KNN imputation, which will all be described below.

- **Mean imputation** - A case of univariate imputation, in which the missing values are replaced with the arithmetic mean of the observed elements. As for categorical data, the missing variable's mode can be used (mode imputation). The simplicity and convenience of this technique makes it the most commonly used [53]. However, mean imputation decreases the variability of the data and distorts its distribution's shape, reducing the standard deviation and measures of association, such as the correlation [58], [68]. In fact, Enders [58] states that this technique leads to biased parameter estimates even under a MCAR mechanism.

- **Regression imputation** - In this approach, the missing values are filled in with predicted data calculated from a regression equation. A different linear regression is built for each incomplete variable based on the remaining ones (multivariate imputation), usually through a complete-case analysis. Since continuous estimates are not suitable for dichotomous variables, a logistic regression can be built instead, as it models the probability of binary outcomes and is able to estimate a replacement value according to that probability [69]. A drawback to regression imputation may arise in datasets with multiple missingness patterns, because each one requires a distinct equation, which complicates the procedure [68]. Furthermore, this technique relies on the assumption of linear relationship between variables, which might be absent. Consequently, the imputation overestimates correlations and attenuates the variability of the data, albeit not as strongly as mean imputation [53], [58].

- **KNN imputation** - This technique is amongst the hot deck imputation methods, in which missing values from an incomplete unit (recipient) are replaced by data extracted from instances similar to the recipient (donors) [70]. The missing attribute in the recipient must be observed in the donors. In KNN imputation, the K closest neighbours, i.e. the K units with the smallest distance to the recipient, are first selected. Then, the missing value is imputed with the average of the K observed values. There are numerous distance measures, such as Minkowski distance, Manhattan Distance, and Euclidean distance, the latter being the most common choice [71]. KNN imputation is applicable to both numeric and non-numeric variables,

maintains the univariate distribution of the data and does not affect its variability to the same extent as other imputation methods [58]. However, the computational time is longer, as it is a slightly more complex technique that requires going through the whole dataset. Moreover, Beretta and Santaniello [72] observed that KNN imputation not only lacks precision when handling attributes with no dependencies in a dataset, but also may introduce spurious associations into the data.

When compared to deletion methods, single imputation has the advantage of maintaining the dataset's size, which permits the use of all units. However, these techniques handle the imputed values as true estimates, distorting standard errors [58], [68].

In MI, on the other hand, various replacement values are computed for each missing data point. Every missing value gets m plausible estimates predicted through the observed data, which generates m imputed datasets. Afterwards, the desired statistical analysis is performed on each imputed dataset and the results are pooled to produce a single set of estimates. Hence, MI comprises three phases: imputation, analysis, and pooling. The last phase may involve averaging parameters estimates, such as regression coefficients, or the use of ensemble techniques, namely bootstrap aggregation and stacking [73].

One of the most popular MI algorithms is Multivariate Imputation by Chained Equations (MICE), proposed by Van Buuren and Groothuis-Oudshoorn [74], which operates under an assumption of MAR missingness. The MICE method starts by replacing all missing values with a temporary "place holder", generally through mean imputation [75]. Afterwards, a permanent imputation is performed in an iterative manner, i.e. addressing one incomplete variable at the time. A regression model is built for a certain incomplete attribute, in which that attribute is the dependent variable and all the other variables are the independent ones. Note that the independent variables are entirely complete, as they contain the "place holders". The missing values in the incomplete attribute are replaced with the predictions from the fitted regression model, and the algorithm moves to the next incomplete attribute. The imputed values in each iteration are used in the subsequent ones, instead of the temporary "place holders". This procedure is repeated for every incomplete attribute, thus completing a cycle. The MICE algorithm can have various cycles, updating the performed imputations in each one. By default, each variable is regressed considering its type: a logistic regression is chosen for binary variables and the predictive mean matching is used for numeric data.

MI overcomes some drawbacks associated with single imputation, particularly the lack of variability. By pooling multiple results, the estimated data will reflect the statistical uncertainty in the imputation process, which results in more precise standard errors [58]. In fact, Peugh and Enders [68] state that this method yields unbiased parameter estimates under a MAR mechanism, although the same may not happen for MNAR data. However, MI may exhibit inferior performances on high dimensional data, as well as data with high MRs [71], [76].

2.4.2.3 Model-based Methods

Model-based methods, as the name suggests, rely on the establishment of a model to represent the complete data, from which inferences will be drawn. This model is related to the data's statistical distribution, whose parameters are commonly estimated through maximum likelihood procedures [52]. Maximum likelihood estimation aims to find the set of population parameter values most likely to have generated the observed data, i.e. that present the optimal fit [68]. An iterative optimisation algorithm, such as the Expectation Maximisation (EM) algorithm, is often used to repeatedly test different sets of parameters until an optimal fit is reached. After this, missing value imputation may be performed through the gathered results.

A first advantage of these methods is the possibility to consider the incompleteness of data when producing estimates of sampling variance [52]. Moreover, maximum likelihood estimation is not only superior to traditional techniques (e.g. mean imputation) under a MCAR mechanism, but also generates unbiased parameter estimates with MAR data [68]. Nevertheless, this method may yield biased parameter estimates under a MNAR mechanism. Furthermore, non-compliance with the assumption of a multivariate normal distribution can bias standard errors, although various corrective procedures for non-normal data have been developed to address this limitation [58]. Lastly, model-based methods present other disadvantages such as a high computational cost and iterative procedures with low speed of convergence [37].

2.4.2.4 Performance Evaluation of Missing Data Handling Approaches

Missing value handling approaches can be evaluated through different criteria, some of which will be described below. More emphasis will be placed on metrics that assess the precision of the imputation process.

1. **Root Mean Squared Error (RMSE)**

   The RMSE represents the standard deviation of the differences between observed values and predicted missing values:

   $$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(o_i - p_i)^2} \tag{2.19}$$

   where o_i denotes the ith observed value, p_i the ith imputed value, and m the number of estimations, i.e. the number of missing values.

   A lower RMSE indicates a narrower difference between true and predicted values, thus evaluating the quality of the imputation procedure. Although being the favoured measurement, RMSE is fairly influenced by outliers [77]. Naturally, this metric can only be computed when an originally complete dataset is available.

2. **Mean Absolute Error (MAE)**

   As for datasets that include dichotomous attributes, the MAE is a more suitable choice [78]. The MAE represents the average difference between observed values and predicted missing values:

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^{m} |o_i - p_i| \tag{2.20}$$

   where o_i denotes the ith observed value, p_i the ith imputed value, and m the number of estimations.

   Therefore, this performance indicator assesses how similar the predicted values are to the real values, with a smaller MAE reflecting a more precise imputation process. For binary variables, this metric represents the proportion of falsely imputed categories, i.e. the error rate, which can also be named MAE for simplicity. This error can also only be calculated when an originally complete dataset is available.

3. **Standard ML Evaluation Metrics**

   After handling the missing data, whether by a complete-case analysis or an imputation method for example, the processed dataset can be given as input to a ML algorithm. Evaluation metrics, such as the accuracy or the AUROC, can assess the impact of each imputation technique on the model's performance. These metrics have been described previously in Section 2.3.2.3.

Chapter 3

Literature Review

This chapter provides a review of literature studies on missing value imputation, and is subdivided into two sections. The first one focuses on state-of-the-art techniques which exploit correlations amongst attributes. In the second section, the application of missing data imputation in the clinical domain is explored and discussed.

3.1 Missing Data Imputation

Missing values are ubiquitous in any real-world dataset, which prompted a growing interest in research projects that address this challenge. The work of Rubin [57] laid the foundation for further studies concerning this issue, paving the way for the development of new approaches. Nevertheless, there are still limitations to be overcome. Throughout this section more focus will be placed on techniques that rely on correlation between values to perform missing data imputation, as such a choice is suggested to be beneficial by several authors.

Regression imputation, where missing data are replaced by predictions from regression equations, is a very common approach [6], [79], [80]. Mishra et al. [79] proposed the Feature Correlation based Missing Data Imputation (FCMI) algorithm, that for each column with missing values (the target) builds a regression model based on other columns that are highly correlated with that specific target. The authors argue that highly correlated attributes capture information about each other. FCMI outperformed KNN and two MI techniques on both numeric and categorical data. However, these experiments were only conducted on MAR features and the results may not generalise to data under the remaining two missingness mechanisms. Moreover, FCMI calculates the correlation coefficients and fits the regression models through a complete-case analysis, and a subset without missing values may not be possible to obtain from a real-world dataset.

Another strategy using regression imputation was presented by Sefidian and Daneshpour [6]. The Correlation Maximization-based Imputation Methods (CMIM) are ten slightly different techniques that attempt to maximise the correlation between the missing attributes and the remaining ones. This paper assumes that the linear correlations within the whole dataset are weaker than those found within certain subsets, and thus data segments with strong correlations are created in order to fit linear regression models. The performed imputation was evaluated under the MAR mechanism, with four MRs. The CMIM approach was superior to competing techniques such as mean imputation, KNN, a DT-based imputation, and an auto-encoder neural network imputation. Nevertheless, CMIM has some drawbacks and the major one pointed out by the authors is the choice of the best technique, among the ten presented, for a given dataset. Since no technique was clearly superior to the others it would be necessary to make a personalised and non-automatic selection for each dataset, which is naturally time-consuming. Furthermore, the optimal parameter values for each imputation method were found through a grid-search strategy, thus increasing the computational cost.

One of the most widely-used approaches is KNN imputation, and variants of this method [7], [78], [81]–[84]. Liu, Lai and Zhang [7] developed the Correlation-based Hierarchical K-Nearest Neighbors (CoHiKNN) algorithm, which utilises the correlation between attributes as weights in the calculation of the distances from each complete record to the target incomplete record. After selecting the nearest neighbours, the missing element is imputed with the average of the observed values. The imputation is phased, as it is performed in segments with an ascending number of missing elements, i.e. it starts with the subset of records with fewest missing values and ends with the ones with the most. The coefficients are initially obtained through a complete-case analysis, and updated in each phase with the union of complete and imputed instances. Although not specified in the paper, the datasets used indicate that experiments were conducted on both numeric and categorical data, under a MCAR mechanism. The CoHiKNN approach was superior when evaluated against mean imputation, a clustering centre-based imputation, and traditional KNN imputation. However, this technique requires a complete subset to compute the first correlation coefficients, which may not be attainable from a real-world dataset, as was stated before.

In order to counter the sparsity of methods specifically addressing the imputation of categorical variables, Faisal and Gerhard [81] proposed a weighted KNN method for nominal data, both binary and multi-categorical. A set of nearest neighbours is obtained for each missing value of every incomplete sample. The developed distance function assigns greater importance, i.e. weight, to covariates highly correlated with the missing attribute. The authors chose Cramér's V as the association measurement. When evaluated on MCAR data, this approach showed a smaller imputation error than methods such as mode imputation, RF-based imputation, and a MI technique.

The so-called data splitting-based techniques have also gained popularity in literature [85]–[88]. Tsai, Miao-Ling and Wei-Chao [85] presented the Class Center based Missing

Value Imputation (CCMVI) method, which starts by partitioning the entire dataset into classes. In the case of a SL problem, the class labels correspond to the outcome values. Each class has a centre, a standard deviation and a threshold based on the euclidean distances between the centre and the complete data samples. The latter is used in the missing data imputation procedure, which was evaluated on categorical, numeric and mixed data, under a MCAR mechanism. However, missing values are only present in the training set, a rather impractical assumption. CCMVI outperformed mean / mode imputation, KNN imputation and SVM imputation for numeric and mixed data types.

A few of these data splitting-based methods require a subset with complete samples to impute the missing values [85], [87], [88], which prompted Bhagat and Singh [86] to propose the Nullify the Missing Values before Imputation (NMVI) approach to overcome this limitation. Similarly to CCMVI, the NMVI method initially splits the entire dataset into classes according to the outcome label of each instance (in a SL problem). Rather than computing the class centres, standard deviations and thresholds using solely complete data, this technique replaces the missing elements with the value zero and uses all class samples, thus excluding the dependency on a complete subset. Experiments were conducted on both numeric and nominal data, with several MRs under all three missingness mechanisms separately applied to the whole dataset. However, the authors do not specify how NMVI handles missing data in the test set for a SL problem, which may indicate that, similarly to CCMVI, this subset is unrealistically assumed to be complete. Performance evaluation showed that NMVI results are on par with techniques such as KNN imputation, CCMVI, and a MI method.

A novel paradigm was proposed by Miao et al. [89] through the development of a mixed interpolation procedure based on feature differences, i.e. a procedure that employs the optimal interpolation technique for each column. Although the term interpolation is used throughout this paper, the given definition is compatible with that of the term imputation. This algorithm dynamically analysed which of the four imputation methods selected by the authors - mean interpolation, regression, SVM, and MI - yielded the best results on every column individually. Experiments showed that the proposed interpolation procedure outperformed all four of the aforementioned imputation techniques on synthetically created missing data, under an unmentioned missingness mechanism. Despite implementing the best method on each attribute, this work does not mention if the proposed procedure is suitable for both numeric and categorical data.

The potential of MI approaches has also been a subject of study [73], [90]–[93]. For instance, Khan and Hoque [90] presented an extension to the MICE algorithm, named Single Center Imputation from Multiple Chained Equation (SICE), that aimed to overcome the complexity of analysing multiple high-dimensional datasets. SICE replaces each numeric (categorical) missing element with the average (mode) of the corresponding values in the differently imputed datasets generated by MICE, thus being described as a hybrid of single and MI. The experimental results showed that the proposed technique generally performed better imputation of binary and numeric data than competitors such

as KNN, SVM, and logistic regression. Even so, the authors did not specify the missingness mechanisms nor MRs on all used datasets, just mentioning that MAR mechanism, with a MR of 10%, was injected onto one of the datasets.

In recent years, there has been a widespread application of Deep Learning (DL) in various fields, and missing value imputation is no exception [94]–[97]. Khan, Wang and Liu [94] proposed the Convolutional Neural Network Imputation (CNNI) approach, which identifies existing correlations within a large dataset and leverages that information to train a convolutional kernel. CNNI performed equal, sometimes better, than KNN imputation, two MI methods and a DL approach, but experiments were carried out only on numerical MCAR data, and thus the results may not generalise.

As for model-based methods, several variants of EM imputation have been developed [78], [98], [99]. Razavi-Far et al. [78] presented the kEMI technique, whose name reflects the implemented combination of the KNN algorithm with EM imputation. For each incomplete record, the former automatically finds the best set of nearest neighbours, and then the latter imputes the missing values by exploring global similarities, i.e. correlations, among the selected donors. kEMI was designed to handle both numeric and categorical attributes, and its performance was superior to other techniques, including standard EM imputation, KNN imputation, and DT-based imputation. Evaluation was conducted under diverse missingness patterns, although the authors do not mention the used mechanism.

Table 3.1 provides an overview of the presented literature studies on missing value imputation. Multiple approaches that exploit correlations between values yield promising results, albeit with recurring limitations such as evaluation for only one missingness mechanism and the requirement for a complete subset. Even though these methods were validated in a controlled environment, i.e. with missing elements synthetically injected into clean datasets, it is essential to account for the variability of a real-world situation before deploying them in such scenario.

Table 3.1: Literature studies on missing value imputation.

Reference	Missing Values	Model Summary	Limitations
Mishra et al. [79]	MAR (MR = 10%)	FCMI: Regression models are built based on highly correlated attributes.	– Only one missingness mechanism was tested. – Needs a complete subset.
Sefidian and Daneshpour [6]	MAR (MR = 30%, 40%, 50%, 60%)	CMIM: Ten techniques that aim to maximise the correlation between the missing features and the remaining ones. Subsets with strong correlations are created to fit regression models.	– Only one missingness mechanism was tested. – Can not directly impute discrete values. – Needs a complete subset.

Continuation of Table 3.1

Reference	Missing Values	Model Summary	Limitations
Liu, Lai and Zhang [7]	MCAR (MR = 5%, 10%, 15%, 20%, 25%)	CoHiKNN: Accounts for the correlation between attributes when selecting the nearest neighbours, which will be used for imputation.	– Only one missingness mechanism was tested. – Needs a complete subset.
Faisal and Gerhard [81]	MCAR (MR = 10%, 20%, 30%)	Weighted KNN approach for nominal data, in which highly correlated features have a greater contribution to the distance.	– Only one missingness mechanism was tested. – Only nominal data was used for the imputation model, although that was the paper's objective.
Tsai, Miao-Ling and Wei-Chao [85]	MCAR (MR = 10%, 20%, 30%, 40%, 50%)	CCMVI: The dataset is split into classes. A threshold for missing value imputation is defined for each class based on the distances between the centre and the samples.	– Only one missingness mechanism was tested. – Missing data is solely in the training set. – Needs a complete subset. – Does not consider correlation between features.
Bhagat and Singh [86]	MCAR, MAR, MNAR (MR = 10%, 20%, 30%, 40%, 50%)	NMVI: Similar to CCMVI, except it replaces the missing values with zero and uses all class samples, thus excluding the dependency on a complete subset.	– Does not specify the handling of missing data in the test set. – Does not consider correlation between features.
Miao et al. [89]	Unknown mechanism; MR = 5%, 10%, 20%, 30%, 40%	Mixed interpolation procedure that employs the optimal imputation technique (mean, regression, SVM or MI) in each column.	– Does not specify the missingness mechanism. – Needs a complete subset. – Does not consider correlation between features. – Does not mention the type of variables for which it is suitable.

Continuation of Table 3.1

Reference	Missing Values	Model Summary	Limitations
Khan and Hoque [90]	MAR (MR = 10%); Information not specified for all used datasets.	SICE: Extension of MICE. Replaces each missing element with the mean / mode of the corresponding values in the differently imputed datasets generated by MICE.	– Only one missingness mechanism was tested. – Does not consider correlation between features.
Khan, Wang and Liu [94]	MCAR (MR = 10%, 20%, 30%, 40%, 50%, 60%, 70%)	CNNI: Identifies existing correlations within a large dataset and leverages that information to train a convolutional kernel.	– Only one missingness mechanism was tested. – Performance evaluated on numeric datasets alone.
Razavi-Far et al. [78]	Unknown mechanism; MR = 1%, 5%, 10%, 20%	kEMI: The KNN algorithm finds a set of donors and EM explores global similarities (correlations) to impute the missing values.	– Does not specify the missingness mechanism, only the pattern. – Needs a complete subset.

3.2 Missing Data Handling in Clinical Records

Similar to the majority of real-world datasets, medical data are inconsistent and incomplete in nature. Missing values must be handled heedfully in order to prevent adverse effects on the performance of any predictive model [52]. Furthermore, ensuring data quality is key when developing a reliable AI-based clinical DSS, since its prediction accuracy might affect the decision-making process [5]. Therefore, efforts have been made towards an efficient processing of missing values.

Firstly, a few comparative studies regarding the performance of well-established and standard imputation approaches on real-world clinical datasets will be presented. These datasets do not have artificially created missing values, but instead present unknown and possibly multiple missingness mechanisms simultaneously. Hence, it is worth assessing the efficiency of such techniques under these less controlled conditions.

Jerez et al. [100] studied the impact of imputation methods on the prognosis accuracy of a real breast cancer problem. The results obtained through a complete-case analysis were compared to those produced by three statistical methods (mean, hot deck imputation, and MI) and three ML-based methods (two DL-based approaches and KNN). Apart from hot deck imputation, all techniques enhanced the prediction accuracy, with KNN leading to the best improvement. Furthermore, the authors stated that the ML-based methods are

a more suitable choice for that particular case study, while noting that this deduction may not generalise to other datasets.

The work conducted by Rahman and Davis [101] yielded similar results for clinical risk prediction on cardiovascular patients. In fact, all ML-based imputation methods chosen for the study (KNN, SVM, DT, and a fuzzy rule-based classification algorithm) exhibited superior performances than the statistical method (mean / mode imputation).

Finally, Austin et al. [92] investigated the performance of MI in clinical research, specifically in a case study concerning mortality prediction on patients hospitalised with heart failure. Although this paper showed that MI outperformed complete-case analysis, it would have been interesting to explore the performance of other imputation techniques in order to draw more relevant conclusions.

On the other hand, some authors developed novel imputation approaches, which are then evaluated on particular clinical case studies [102]–[106]. A brief overview of a few techniques will be provided, seeking to identify and discuss their limitations and successful strategies.

Madhu et al. [102] proposed the missXGBoost method, which trains an Extreme Gradient Boosting algorithm to impute continuous values in both continuous and discrete attributes. This technique yielded a superior classification accuracy than a few competitors when tested on several benchmarks medical datasets. As for research works with a specific clinical objective, Jaques et al. [103] designed the Multimodal Autoencoder, a DL method that handles missing data within a real mood prediction problem. Since this model was trained considering missing data, its performance is not affected as much as in techniques that did not take this into account. However, neither Madhu et al. [102] nor Jaques et al. [103] considered the correlations between features, which combined with their requirement for a complete subset, might avert its employment in other clinical contexts.

In the study of Alzheimer's disease, Tabarestani et al. [105] developed a multitask learning method for progression prediction. The authors stated that this approach is robust to the negative effects of missing data, as it captures the dependencies between different feature representations and preserves enough information to ensure a fairly high performance accuracy. Yoon, Zame and van der Schaar [106] proposed a DL model which also exploited the relationship among features, particularly the correlation within and across data streams. This work was not within the scope of a specific clinical case study, but was tested on several real-world medical datasets and demonstrated considerable improvements over state-of-the-art techniques, including MICE and DL-based approaches.

Despite a noticeable research endeavour to overcome the challenges posed by missing values in clinical data, complete–case analysis is still one of the most common strategies to address this issue [92], [107]. Albeit straightforward, this approach regards missing values as a disposable nuisance, overlooking the reason for their occurrence. Carrying out an inattentive complete-case analysis may have deleterious consequences, as some authors assert that missingness can be informative [55], [108].

Moreover, missing data imputation methods often present limitations such as the necessity for a complete subset and validation for numeric variables alone. In the health care domain, information is frequently collected by multiple sources and a reliable DSS must be able to handle the inherent variability of real-world clinical datasets. Additionally, professionals' trust in a AI-based DSS increases if it is clear how the model reached its decisions and what inputs were used for the prediction. Although DL models might outperform traditional techniques, such as KNN and SVM, they may be considered "black boxes" in terms of their inner-working. Even if methods for interpreting these DL models are employed, their complexity and the need for a large amount of data may hinder their use in clinical decision-making [109], [110]. Similarly, an imputation procedure suitable for a real health care scenario ought to have an optimised performance, whilst maintaining an adequate level of interpretability.

To the best of our knowledge, exploration of the concept of correlation in medical research is rather scarce, whether in novel imputation approaches [89], [102]–[104], whether in frameworks that make use of existing standard methods [5], [82]. In fact, most ML applications consider highly correlated features as redundant, discarding one of them before further processing. However, this procedure is not advisable for an imputation model, as various studies have shown the beneficial effects of taking correlation into account [6], [7], [78], [79], [81], [94], [105], [106]. Hence, it is worth investigating the potential contribution of correlation for the imputation of missing medical data.

Chapter 4

Datasets

This chapter introduces the five datasets used throughout this dissertation. Firstly, three publicly available and complete datasets were chosen and manipulated to enable a comprehensive study that included various MRs of the three missingness mechanisms. Then, since the main focus of this work was to study missing value imputation in clinical contexts, two real-world medical datasets were selected and described. The provided characterisation focuses on the types of variables and the distribution of missing values within every dataset.

4.1 UCI Machine Learning Repository Datasets

In order to evaluate missing data imputation approaches and compare them against state-of-the-art methods, three complete and publicly available datasets were selected from the UCI Machine Learning Repository [111]. This collection of databases was created by the Center for Machine Learning and Intelligent Systems at the University of California, Irvine (UCI). The selection process was primarily based on the type of variables of each dataset, as it is essential to ensure that these methods perform a suitable and efficient imputation regardless of the attribute's type.

Within this work, multiclass nominal variables were one-hot encoded, which resulted in the creation of several binary attributes associated with each original nominal variable. Furthermore, ordinal encoding was applied to ordinal attributes, which are then treated as numeric variables, a very common and often reasonable assumption [112]. Hence, it is only relevant to assess the imputation accuracy on both numeric and binary variables.

Taking this into consideration, three complete datasets were retrieved from the aforementioned repository: **Wine Data Set**, only comprising numeric variables; **SPECT Heart Data Set**, which solely includes binary attributes; **Statlog (Heart) Data Set**, a mixed-type

dataset. The categorical variables within the latter were encoded as mentioned above. An overview of these datasets (after encoding) is provided in Table 4.1.

Table 4.1: Overview of the datasets retrieved from the UCI Machine Learning Repository.

Name	Number of instances	Number of attributes	Variable types	Task
Wine Data Set	178	13	Numeric	Multiclass Classification
SPECT Heart Data Set	187	22	Binary	Binary Classification
Statlog (Heart) Data Set	270	22	Numeric and Binary	Binary Classification

Synthetic missing values were injected into these three datasets, under all three missingness mechanisms with three different MRs (10%, 30%, and 50%), totaling nine synthetic datasets per each UCI Machine Learning Repository dataset. The adopted procedures are further described in Section 5.2.

4.1.1 Wine Data Set

The Wine Data Set was obtained through a chemical analysis of three different types of wines. It contains 178 samples, which can be classified into three categories (wine types) based on 13 attributes. These attributes, whose names are displayed in the first column of Table 4.2, are the numeric quantity of chemical constituents found in the wine. Table 4.2 also provides an overview of each of the nine datasets that were generated through the injection of missing values, detailing which variables are missing under the corresponding mechanism and MR.

Table 4.2: Characterisation of the missingness injected into the Wine Data Set. The crosses indicate which variables are missing in each synthetic dataset. Total missingness is the ratio between the number of missing values and the total number of values. Incomplete samples is the proportion of instances with at least one missing attribute.

	MCAR			MAR			MNAR		
	10%	30%	50%	10%	30%	50%	10%	30%	50%
Numeric									
Alcohol	X	X	X						
Malic acid	X	X	X						
Ash	X	X	X						
Alcalinity of ash	X	X	X	X	X	X	X	X	X
Magnesium	X	X	X						
Total phenols	X	X	X						
Flavanoids	X	X	X	X	X	X	X	X	X
Nonflavanoid phenols	X	X	X						

Continuation of Table 4.2

	MCAR			MAR			MNAR		
	10%	30%	50%	10%	30%	50%	10%	30%	50%
Proanthocyanins	X	X	X	X	X	X	X	X	X
Color intensity	X	X	X	X	X	X	X	X	X
Hue	X	X	X						
OD280/OD315 of diluted wines	X	X	X	X	X	X	X	X	X
Proline	X	X	X	X	X	X	X	X	X
Total Missingness	10%	30%	50%	5%	14%	23%	5%	14%	23%
Incomplete Samples	74%	100%	100%	47%	84%	97%	46%	88%	99%

4.1.2 SPECT Heart Data Set

The SPECT Heart Data Set results from cardiac Single Proton Emission Computed Tomography (SPECT) images, which can be classified into normal or abnormal. The available training set includes instances from 187 patients with 22 binary attributes extracted from the images, whose names are shown in the first column of Table 4.3. Similarly to the previous example, Table 4.3 presents which variables are missing in each of the nine synthetic datasets.

Table 4.3: Characterisation of the missingness injected into the SPECT Heart Data Set. The crosses indicate which variables are missing in each synthetic dataset. Total missingness is the ratio between the number of missing values and the total number of values. Incomplete samples is the proportion of instances with at least one missing attribute.

	MCAR			MAR			MNAR		
	10%	30%	50%	10%	30%	50%	10%	30%	50%
Binary									
F1	X	X	X						
F2	X	X	X						
F3	X	X	X						
F4	X	X	X						
F5	X	X	X	X	X	X	X	X	X
F6	X	X	X						
F7	X	X	X						
F8	X	X	X						
F9	X	X	X						
F10	X	X	X						
F11	X	X	X	X	X	X	X	X	X
F12	X	X	X	X	X	X	X	X	X
F13	X	X	X	X	X	X	X	X	X

Continuation of Table 4.3

	MCAR			MAR			MNAR		
	10%	30%	50%	10%	30%	50%	10%	30%	50%
F14	X	X	X						
F15	X	X	X	X	X	X	X	X	X
F16	X	X	X						
F17	X	X	X						
F18	X	X	X	X	X	X	X	X	X
F19	X	X	X	X	X	X	X	X	X
F20	X	X	X	X	X	X	X	X	X
F21	X	X	X	X	X	X	X	X	X
F22	X	X	X	X	X	X	X	X	X
Total Missingness	10%	30%	50%	5%	15%	25%	5%	15%	25%
Incomplete Samples	92%	100%	100%	23%	61%	96%	20%	64%	96%

4.1.3 Statlog (Heart) Data Set

The Statlog (Heart) Data Set is used to assess the presence of heart disease in a binary classification problem. It encompasses cardiac data from 270 patients, summarised into 13 numeric and categorical attributes. After performing the previously described categorical encoding, the processed dataset has 6 numeric and 16 binary variables, resulting in the 22 attributes exhibited in the first column of Table 4.4. As before, Table 4.4 depicts which variables are missing in every synthetic dataset.

Table 4.4: Characterisation of the missingness injected into the Statlog (Heart) Data Set. The crosses indicate which variables are missing in each synthetic dataset. Total missingness is the ratio between the number of missing values and the total number of values. Incomplete samples is the proportion of instances with at least one missing attribute.

	MCAR			MAR			MNAR		
	10%	30%	50%	10%	30%	50%	10%	30%	50%
Numeric									
Age	X	X	X						
Resting BP	X	X	X	X	X	X	X	X	X
Serum cholesterol	X	X	X	X	X	X	X	X	X
Maximum HR achieved	X	X	X	X	X	X	X	X	X
Oldpeak	X	X	X						
Major vessels	X	X	X	X	X	X	X	X	X
Binary									
Sex	X	X	X						

Continuation of Table 4.4

	MCAR			MAR			MNAR		
	10%	30%	50%	10%	30%	50%	10%	30%	50%
Chest pain_1	X	X	X						
Chest pain_2	X	X	X						
Chest pain_3	X	X	X						
Chest pain_4	X	X	X	X	X	X	X	X	X
Fasting blood sugar	X	X	X						
Resting EKG_0	X	X	X						
Resting EKG_1	X	X	X						
Resting EKG_2	X	X	X	X	X	X	X	X	X
Exercise induced angina	X	X	X	X	X	X	X	X	X
Slope_1	X	X	X						
Slope_2	X	X	X	X	X	X	X	X	X
Slope_3	X	X	X	X	X	X	X	X	X
Thal_3	X	X	X						
Thal_6	X	X	X	X	X	X	X	X	X
Thal_7	X	X	X	X	X	X	X	X	X
Total Missingness	10%	30%	50%	5%	15%	25%	5%	15%	25%
Incomplete Samples	90%	100%	100%	33%	83%	99%	44%	81%	100%

4.2 Osteoporosis Dataset

The Osteoporosis Dataset assembles publicly available data from the 2013-2014 cycle of the National Health and Nutrition Examination Survey (NHANES) [113]. The NHANES is a research program that aims to monitor the health and nutritional status of adults and children in the United States. The survey participants are randomly selected and, therefore, can present a variety of health statuses.

This study combines interviews and physical examinations in order to assess the prevalence of diseases and medical conditions, such as anaemia, diabetes, osteoporosis and many more. The NHANES interviews, conducted at the respondents' homes, collect demographic, socioeconomic, dietary and health-related information. The physical examinations are performed by trained medical personnel, and include laboratory tests and physiological measurements.

The working dataset consists of 37 variables and 1643 subjects, which can be classified into 3 conditions: normal, osteopenia, and osteoporosis. Osteoporosis is a skeletal disorder characterised by a deterioration of the bone architecture and a reduction of bone mass, which leads to increased bone fragility and susceptibility to fracture [114]. Osteopenia is

the condition that precedes osteoporosis. Subjects with osteopenia have a bone mineral density lower than the reference values for a healthy individual, but not low enough to be diagnosed with osteoporosis. The osteopenia and osteoporosis classes were combined into a single one, thereby transforming this study into a binary classification problem. As for the collected data, Table 4.5 gives a brief characterisation of each feature group within the Osteoporosis Dataset, including the number and type of variables, and the average MR. Table A.1 provides a more detailed description of each variable.

Table 4.5: Brief characterisation of the Osteoporosis Dataset.

Feature Group	Attributes	Average MR
Demographics	1 numeric, 1 nominal and 2 ordinal	(1.1 ± 1.9)%
Nutrition	6 numeric	(7.6 ± 0.2)%
Blood pressure	2 numeric	(3.3 ± 0.3)%
Anthropometrics	2 numeric	(0.6 ± 0.1)%
Physical fitness	1 numeric	10.0%
Blood lipids	4 numeric	(27.8 ± 25.4)%
Hormones	3 numeric	(8.9 ± 4.5)%
Biochemistry	2 numeric	(3.1 ± 0.7)%
Physical activity	11 numeric	(10.0 ± 0.1)%
Lifestyle	2 numeric	(9.8 ± 0.0)%

The variables DMDEDUC2 and INDFMIN2, reported in Table A.1, had two additional categories, "Dont' Know" and "Refused", which were considered missing values in this work. Recalling the definition of missing data given by Little and Rubin [52], this is a reasonable assumption since these responses hide values that would be meaningful for analysis if observed.

In summary, the MRs in each attribute vary from 0.0% to 53.4%, and the dataset has a proportion of 72.0% incomplete samples, i.e. subjects with at least one missing value. Furthermore, 38.8% of the individuals are classified as normal (healthy) and the remaining 61.2% were diagnosed with either osteoporosis or osteopenia. The average age of all participants is 58 ± 12 years, with those classified as normal having a mean of 58 ± 10 years and the remaining, considered not healthy, an average age of 62 ± 12 years.

After performing categorical encoding, the final working dataset was left with 43 variables, 36 of which are numerical and 7 are binary. A special encoding had to be applied to the variable INDFMIN2: although it has 12 ordered categories, the remaining 2, "$20,000 and Over" and "Under $20,000", have binary characteristics that do not fit into the variable's ranking. Hence, the 12 ordered categories were ordinally encoded, while the other 2 were considered missing values. Then, an additional attribute named INDFMIN2_binary was created, in which the values of 0 and 1 were attributed, respectively, to the categories that represented range values under $20,000, and $20,000 and over. As a consequence of this procedure, the proportion of incomplete samples rose to 73.2%.

4.3 Cardiothoracic Surgery Dataset

The Cardiothoracic Surgery Dataset contains clinical and demographic information retrieved by the Cardiothoracic Surgery Service of Hospital de Santa Marta in Portugal, from 2011 to 2019. For each subject, the collection started during the pre-surgery period and extended up to one year after the surgical procedure.

Furthermore, measurements were carried out in compliance with the data fields found in the Adult Cardiac Database, from the Quality Improvement Programme (QUIP) [115]. The QUIP, created by the European Association for Cardio-Thoracic Surgery (EACTS), establishes a set of guidelines to unify cardiac surgical databases, aiming to motivate the improvement of clinical outcomes.

The dataset contains records from 8122 patients and was used to predict the occurrence of complications within three months after hospital discharge, in a binary classification problem. Table 4.6 displays the various incidents considered in the post-discharge complications, as well as the assigned label. The outcome variable takes the value of 1 when an incident labelled as "Severe" or "Death" occurs, and the value of 0 otherwise. Since a patient can experience more than one complication, the outcome variable refers to the most severe incident suffered by each subject within three months after discharge. Patients that did not experience a single complication have an outcome of 0. Subjects in which only the incident labelled as "Unknown" was reported were excluded from the analysis.

Table 4.6: Incidents and labels of the Cardiothoracic Surgery Dataset's outcome.

Incident	Label	Outcome
Impossible do reach	Unknown	-
Superficial infection	Light	0
Complication not requiring admission to the ICU		
Atelectasis / pneumonia / pleural effusion / pneumothorax		
Fibrillation or arrhythmia requiring treatment		
Complication requiring surgery (cardiac or other)	Severe	1
Complication requiring admission to the ICU		
Stroke with deficits persisting for >72 hours		
Infection/instability of the sternum		
Angioplasty		
Endocarditis and/or sepsis		
Definitive pacemaker		
Infarction		
Acute respiratory distress syndrome		
Dialysis or haemofiltration		
Sudden death	Death	

Data was collected on 106 medically relevant variables, which can be grouped into the categories shown in Table 4.7. As before, Table 4.7 provides a brief characterisation of each feature group, including the number and type of variables, and the average MR. Table A.2 gives a more thorough description of the Cardiothoracic Surgery Dataset.

Table 4.7: Brief characterisation of the Cardiothoracic Surgery Dataset.

Feature Group	Attributes	Average MR
Hospitalisation	4 dates	(23.7 ± 47.2)%
Cardiac history	4 ordinal and 1 binary	(16.1 ± 35.3)%
Previous interventions	2 dates, 1 ordinal and 1 nominal	(44.9 ± 51.6)%
Pre-operative risk factors	4 numeric, 3 ordinal, 4 nominal and 4 binary	(0.9 ± 0.2)%
Pre-operative haemodynamics & catheterisation	1 date, 5 numeric, 2 ordinal, 1 nominal and 1 binary	(49.3 ± 42.8)%
Pre-operative status & support	4 binary	(0.6 ± 0.2)%
Operation	1 text, 2 ordinal and 4 nominal	(27.2 ± 46.1)%
Coronary surgery	2 numeric and 2 nominal	(61.4 ± 0.7)%
Valve surgery	1 code, 8 numeric, 3 ordinal, 10 nominal and 7 binary	34.8 ± 30.2)%
Cardiac Surgery Morbidity Scale	1 numeric and 8 nominal	(7.6 ± 22.1)%
Discharge details	2 nominal and 1 binary	(32.4 ± 55.6)%
Patient demographics and autocalculations	1 code, 10 numeric, 1 nominal and 1 binary	(7.8 ± 26.4)%

The MRs in each attribute vary from 0.0% to 99.9%, and the dataset has a proportion of 100.0% incomplete samples, i.e. every individual has at least one missing value. An initial pre-processing was performed, which involved the removal of patients that died before being discharged, replacement of missing values based on domain knowledge, and binarization of features with high cardinality, such as dates. This pre-processing also aggregated in-hospital complications and lengths of hospital stay into new features, and removed variables updated after hospital discharge, e.g. patient status and date of death.

The final working dataset has 5625 subjects. After applying categorical encoding, the resulting number of features is 119, of which 31 are numeric and 88 are binary. Contrary to the previous cases, the ordinal attributes were both ordinally and one-hot encoded in order to investigate if each category could be useful for imputation as an individual feature, instead of solely resorting to the ordinal variable from a continuous perspective.

After the initial pre-processing, the MRs in each attribute range from 0.0% to 82.4%, and 94.1% of the instances are incomplete. Moreover, 92.7% of the patients had light or no complications three months after discharge, and the remaining 7.3% either experienced severe complications or died. From the first group (outcome value is 0), 38.9% are female and have an average age of 65 ± 13 years, whereas 43.4% of the subjects from the second group are female and have an average age of 68 ± 12 years.

In order to perform a time-based analysis, patients whose information was collected in 2019 were assembled in a separate test subset. This group contains 667 subjects, from which 92.7% reported light or no complications three months after discharge. Therefore, the distribution of class labels is maintained in this test set.

Chapter 5

Methodologies

This chapter includes a comprehensive description of the methodologies adopted to answer the RQs raised in Section 1.2. A general approach is first presented, followed by a detailed characterisation of each step within the outlined framework.

5.1 General Approach

This dissertation aims to address the challenges imposed by incomplete real-world datasets on the performance of ML models. To this end, the concept of correlation will be studied and later integrated within novel missing data imputation strategies.

Following a classical approach, the first step of this work involved the selection of case studies, i.e. datasets to perform experiments on, thus allowing the validation of the proposed methods. Three complete datasets and two real-world medical databases were picked, and a description for each one is presented in Chapter 4. The selected databases had a verified objective that guaranteed the credibility of the study while also avoiding the redundancy generated by retrieving some definition from the data.

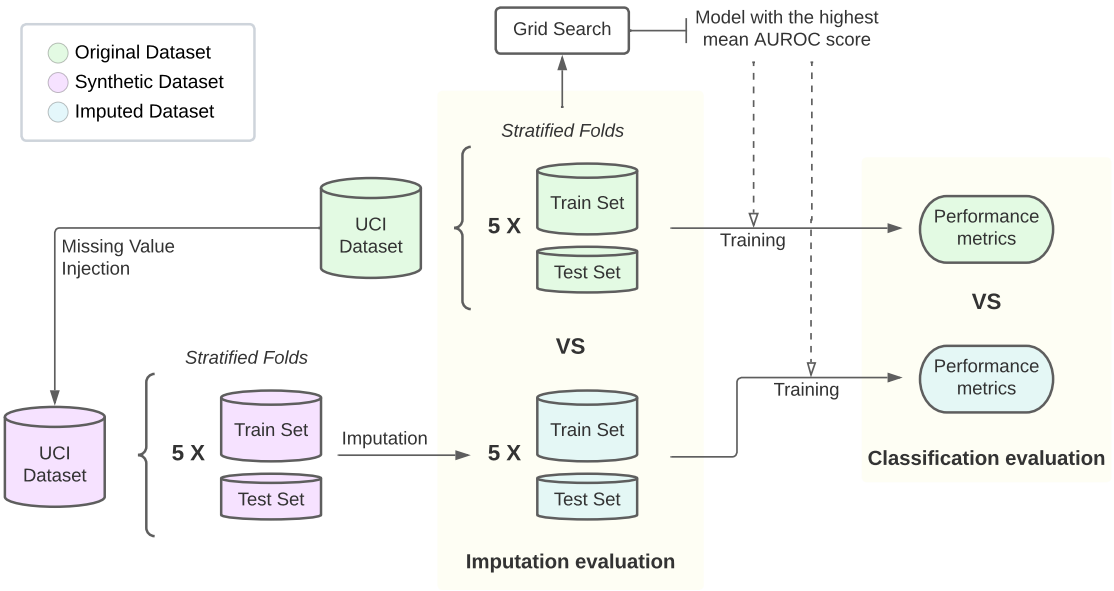
In order to evaluate the performance of several imputation techniques under different missingness mechanisms, and with various MRs, synthetic missing values were injected into the three complete datasets from the UCI Machine Learning Repository. Section 5.2 outlines the adopted procedures. The two real-world medical datasets were not manipulated in this manner, nor is it advised to do so, as they naturally present missing values under unknown and possibly multiple missingness mechanisms simultaneously.

Then, taking inspiration from several works that pointed out the benefits of accounting for correlation when imputing missing values, a correlational study was conducted. In addition to the correlation between the values of two variables, which is the most traditional measurement, two other relationships were assessed: the correlation between the values

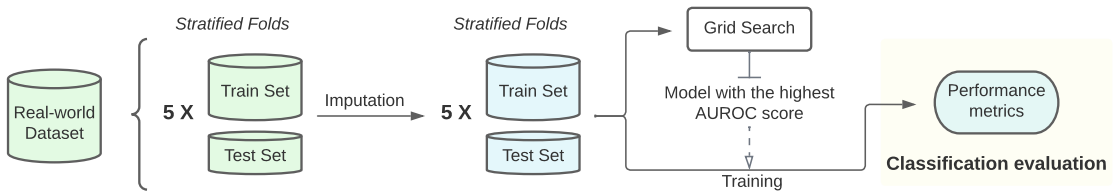
of a variable and the missingness pattern of every other, and the correlation between the missingness patterns of two distinct attributes. Section 5.3 provides a description of this study and its objectives.

The subsequent step consisted of implementing various existent imputation methods, both traditional and state-of-the-art. A comparative evaluation enabled the identification of limitations and strengths associated with each technique, complementing the information found in the literature.

Furthermore, three novel correlation-based imputation approaches were proposed and evaluated against the implemented techniques. The final step consisted precisely in this performance evaluation, which can be subdivided into two types: imputation evaluation and classification evaluation. The former is only available for the three UCI Machine Learning Repository datasets, as it assesses the quality of the imputation procedure by comparing the imputed values with the original ones (ground truth). Figure 5.1 depicts a schematic representation of the followed framework for a final performance evaluation. The illustrated steps are further described in the succeeding sections.



(a) UCI Machine Learning Repository datasets



(b) Real-world datasets

Figure 5.1: Performance evaluation of the missing data imputation models in (a) the UCI Machine Learning Repository datasets, and (b) Real-world medical datasets.

5.2 Injection of Synthetic Missing Data

As aforementioned, artificial missing values were created within the three UCI Machine Learning Repository datasets. For this end, the package missMethods [116] from the Comprehensive R Archive Network (CRAN) was used, as it supplies functions for the injection of missing data. CRAN is a repository of code and documentation for R, and thus the necessary programming for this task was carried out using this language.

The package missMethods is inspired in the work of Santos et al. [117], offering the possibility to implement nearly all the approaches to synthetic missing data generation compiled by the authors. Three functions from this package were selected, each one implementing a different missingness mechanism in a multivariate configuration. The algorithms that originated these functions will be described below.

The $MCAR1_{unifo}$ approach, proposed by Twala [118], was chosen for the MCAR mechanism. The subscript *unifo* was used by Santos et al. [117] to denote multivariate configurations. Within this particular algorithm, the missing elements are obtained through Bernoulli trials: for each observation of every variable, the binary outcome of a Bernoulli trial determines whether the value will be removed or not. The probability of success in that Bernoulli trial is the MR, which will be equal in all the features. The respective implementation in the package missMethods receives this probability of missingness as an argument and has a parameter to enforce a deterministic number of missing values.

As for the MAR mechanism, a combination of the $MAR1_{unifo}$ and $MAR2_{unifo}$ algorithms, proposed respectively by Gariarena and Santana [119] and Twala [118], was performed. Firstly and following Twala [118], several pairs of highly correlated features are created, in which one of them determines missingness in the other. For an odd number of variables, a single triple is formed. In each pair (or triple), the determining feature is the most correlated with the outcome variable. The missing elements on the non-determining feature correspond to the positions of the k lowest values of the determining feature. Instead of defining a MR for the entire dataset as performed by Twala [118], the strategy followed by Gariarena and Santana [119] was prefered where each missing column has an equal MR. The corresponding function in the package missMethods receives, among other parameters, the features in which the missing values will be created and the probability of missingness, i.e. the MR, which then defines the value of k.

Lastly, the $MNAR1_{unifo}$ and $MNAR2_{unifo}$ algorithms, based on the works of Gariarena and Santana [119] and Twala [118] respectively, were merged to generate MNAR data. Similar to the previous procedure, pairs of highly correlated features are formed in which only one of the variables are injected with synthetic missing values. For triples, two variables have missing data. In compliance with the $MNAR2_{unifo}$ method, the lower values of each non-determinant feature are then deleted. However, the amount of missing elements per variable is determined by the MR, equal in all missing features, as proposed by Gariarena and Santana [119]. The missMethods' function that implements this approach also accepts as input the name of the missing variables, as well as the MR.

Table 5.1 displays the implemented missMethods' functions and non-default parameters, along with their corresponding algorithm from Santos et al. [117].

Table 5.1: Implemented missMethods' functions and their respective algorithm from Santos et al. [117]. Only the parameters with a non-default value are specified.

Santos et al. [117] Algorithm (Function)	**Parameters**
$MCAR1_{unifo}$ (delete_MCAR)	**p** $\in \{0.10, 0.30, 0.50\}$
$MAR1_{unifo}$ and $MAR2_{unifo}$ (delete_MAR_censoring)	**p** $\in \{0.10, 0.30, 0.50\}$ Wine Data Set: **cols_mis** = ['Flavanoids', 'OD280/OD315 of diluted wines', 'Alcalinity of ash', 'Proanthocyanins', 'Color intensity', 'Proline'] **cols_ctrl** = ['Total phenols', 'Hue', 'Ash', 'Nonflavanoid phenols', 'Malic acid', 'Magnesium'] SPECT Heart Data Set: **cols_mis** = ['F5', 'F11', 'F12', 'F9', 'F13', 'F20', 'F21', 'F19', 'F15', 'F18', 'F22'] **cols_ctrl** = ['F1', 'F6', 'F7', 'F4', 'F8', 'F16', 'F3', 'F10', 'F14', 'F17', 'F2'] Statlog (Heart) Data Set: **cols_mis** = ['Resting EKG_2', 'Thal_7', 'Slope_2', 'Chest pain_4', 'Maximum HR achieved', 'Slope_3', 'Exercise induced angina', 'Serum cholesterol', 'Resting BP', 'Major vessels', 'Thal_6'] **cols_ctrl** = ['Resting EKG_0', 'Thal_3', 'Slope_1', 'Chest pain_3', 'Age', 'Oldpeak', 'Chest pain_2', 'Sex', 'Chest pain_1', 'Fasting blood sugar', 'Resting EKG_1']
$MNAR1_{unifo}$ and $MNAR2_{unifo}$ (delete_MNAR_censoring)	Same parameters as delete_MAR_censoring

Missing values were injected into the UCI Machine Learning Repository datasets under all three missingness mechanisms individually, with three different MRs: 10%, 30%, and 50%. Therefore, nine synthetic datasets were created for each of these three databases, and a brief characterisation of their missingness can be found in Tables 4.2, 4.3 and 4.4 from Chapter 4.

5.3 Correlational Study

The literature review presented in Chapter 3 compiled several studies showing that correlation had a beneficial impact on the imputation of missing values. In order to further comprehend the vastness of the information provided by this form of dependency, a correlational study, i.e. an exploratory analysis on the relations between two variables, was conducted. This study was divided into two stages.

5.3.1 First Stage

The first stage focused on the different correlation measurements. As discussed in Section 2.2.1.2, there are a myriad of coefficients, each suitable for specific types of variables. Thus, a brief comparison was carried out between coefficients that assess the correlation among the same types of variables, i.e. coefficients that lie in the same cell in Table 2.1. The main objective of this first stage was to select the correlation coefficients that would be used by the proposed imputation methods.

Considering the type of variables present in the UCI Machine Learning Repository datasets, two distinct comparative assessments were performed. The first was between coefficients that measure the correlation among two numeric attributes, such as Pearson's coefficient, Spearman's rank coefficient, and Kendall's Tau coefficient. The second was between coefficients that measure the correlation among two binary attributes, i.e. the Phi coefficient and the Cramér's V. In this stage, the correlation between a numeric and a binary attribute was not addressed because there are no alternatives to the point biserial coefficient, which was then automatically selected to measure this type of relationship in subsequent steps. The performed analyses were solely carried out on the three UCI Machine Learning Repository datasets.

The module scipy.stats from Python's library SciPy provided functions to calculate all the necessary coefficients except for Cramér's V, which was implemented using Python. After this stage, one correlation measurement was selected per relationship type (numeric-numeric, numeric-binary, and binary-binary), resulting in a total of three coefficients.

5.3.2 Second Stage

The second stage of this study encompassed a correlational analysis of every missingness mechanism, using the datasets with injected missing values generated from the three UCI Machine Learning Repository databases. As previously stated, these datasets were chosen based on their type of variables, which permits a more complete assessment. This stage investigated how correlation captures the relationships associated with each missingness mechanism and sought to identify distinctive traits between them.

Firstly, the effect of the missingness mechanism and MR on the correlation between variables was evaluated. The correlation matrix for each dataset with injected missing values was compared with the correlation matrix of the corresponding original dataset.

Furthermore, since the missingness mechanisms describe generic relations between data and missingness, it is interesting to investigate whether correlation can capture these relationships. For example, MAR missingness depends solely on the values of another attribute, which may indicate that the missingness pattern of a MAR variable has a strong correlation with the values from another attribute.

In order to study this correlation between values and missingness patterns within a dataset, a missingness indicator matrix M was first obtained. Let f denote the total number of features and $f_{\text{miss}} \leq f$ the number of features with missing values. f_j represents the

jth feature, where $j \in \{1, 2, ..., f\}$, and $f_{\text{miss},i}$ represents the ith incomplete feature, where $i \in \{1, 2, ..., f_{\text{miss}}\}$. Furthermore, consider that C_{vm} is a $f_{\text{miss}} \times f$ matrix that will store the computed correlations. For every pair $\{i, j\}$, with $f_{\text{miss},i} \neq f_j$, $C_{\text{vm}}[i, j]$ stores the correlation between the values of f_j and the missingness pattern of $f_{\text{miss},i}$, its binary missingness indicator. Matrix C_{vm}, or *var-miss* matrix, was computed for all datasets with injected missing values, under all missingness mechanisms. Then, the matrices were examined to check whether they clearly captured strong relationships, keeping in mind that the variables that determined the missingness in the MAR scenario were known.

In addition to the previous assessment, the correlation between two missingness patterns, or nullity correlation, was also investigated for every missingness mechanism, seeking to identify if any relevant relation existed. For instance, a negative high correlation between two missingness patterns may indicate that the corresponding variables are almost never observed simultaneously.

5.4 Data Splitting and Initial Pre-processing

From a predictive modelling perspective, it is fundamental to first split the working dataset into train and test subsets prior to commencing missing data imputation, or any other pre-processing procedure, in order to prevent data leakage. Within this dissertation, encoding, data scaling and missing data handling were initially performed on the train set, and the knowledge obtained was leveraged to transform the test set.

The data splitting process ought to account for the nature of the case study, as well as the characteristics of the working dataset. In this particular work, it is important to consider that medical databases frequently present unbalanced class labels. For instance, 92.7% of the samples from the Cardiothoracic Surgery Dataset are negative cases and only 7.3% are from the positive class, as noted in Section 4.3.

A stratified 5-fold strategy was applied in order to preserve the proportion of instances for each class target in both train and test subsets. Hence, each working dataset was divided into five different train and test sets with equal distributions of samples per class label, which secured a fair performance evaluation. Note that in the Cardiothoracic Surgery Dataset a grouped stratified 5-fold strategy had to be adopted instead. This approach groups the instances by patient, thus ensuring that the same subject is not represented in both the training and test subset, which could lead to performance overestimation. It was not necessary to take this precaution in the Osteoporosis Dataset as it was previously ascertained that there were no separate instances belonging to the same patient.

As illustrated in Figure 5.1(a), the UCI Machine Learning Repository datasets were only divided into train and test subsets after the injection of missing values since some of the missingness mechanisms involved in this manipulation depend on parameters found from the full dataset. Consider the case where a variable has missing elements when its value is below a threshold equal to a fraction of the maximum value. The maximum value of the training set could not be used, as nothing precludes a higher value from appearing

in the test subset. Moreover, generating missing values only in the training set is not an option, as it would make the problem unrealistic. Consequently, this course of action was pursued, stressing that no data leakage is believed to occur because the true values of the missing elements are not used nor provide information to the ML model.

Afterwards, the initial pre-processing consisting of categorical encoding and data scaling was carried out firstly within each training set. Chapter 4 specified the encoding performed in every dataset. As for data scaling, a normalisation to the range $[0, 1]$ was applied to all features. The normalisation of both train and test subsets of the UCI Machine Learning Repository datasets with injected missing values was based on the training set of the corresponding original database. This ensures that the observed values in each synthetically generated dataset are equal in the respective complete dataset, allowing a valid comparison between the predicted (imputed) values and the ground-truth.

5.5 Missing Data Imputation

This section encompasses the major contribution of this dissertation, i.e. the development of three novel correlation-based imputation approaches, which were evaluated against existing techniques. Before introducing these three approaches, the methods from the literature used as baseline for comparison are presented.

5.5.1 Traditional and State-of-the-art Imputation Methods

Various existent imputation methods, both traditional and state-of-the-art, were implemented. A comparative study was conducted, aiming to identify limitations and strengths associated with each technique, complementing the information found in the literature. Furthermore, these methods served as benchmarks, permitting a more informative evaluation of the proposed imputation techniques.

Overall, seven imputation methods were selected: Mean / Mode, Regression, KNN, CMIM [6], CoHiKNN [7], NMVI [86], and MICE [74]. This selection sought to encompass both common and modern techniques with diverse baseline strategies. Mean / Mode imputation, Regression, KNN, and MICE were described in Section 2.4.2.2, whereas CMIM, CoHiKNN and NMVI were briefly discussed in Section 3.1. Table 5.2 shows all the parameter values of interest that were tested for each method.

A replication of the imputation performed by CMIM, CoHiKNN and NMVI was implemented based on the interpretation of the corresponding papers, as no existing implementation published by the authors was found. Thus, the results obtained may be different from those reported in the literature. Since NMVI did not specify how it handles missing data in the test set it was necessary to make the assumption that each sample in this subset is assigned to the class whose centre is closest to it (using euclidean distance).

Additionally, regression imputation was also implemented from scratch. For the remaining techniques, the classes SimpleImputer and KNNImputer from Python's library

scikit-learn [120], and the package mice [74] from CRAN were used. The necessary programming for this task was carried out using Python and R.

Table 5.2: Parameter values for the selected traditional and state-of-the-art methods.

Method	Parameters
Mean / Mode imputation	**missing_values** = np.nan
Regression imputation	-
KNN	**missing_values** = np.nan, **n_neighbors** $\in \{5, 10, 15\}$, **weights** = 'distance', **metric** = 'nan_euclidean'
CMIM	**percentage** $\in \{0.1, 0.5, 0.9\}$, **threshold** $\in \{0.1, 0.5, 0.9\}$
CoHiKNN	**n_neighbors** $\in \{5, 10, 15\}$
NMVI	-
MICE	**defaultMethod** = c("pmm", "logreg", "polyreg", "polr"), **m** = 3, **maxit** = 10, **seed** = 42

5.5.2 Proposed Imputation Methods

In order to contribute to biomedical research surrounding missing values, three novel correlation-based imputation techniques have been developed: Correlation Weighted K-Nearest Neighbour Imputation (CWKNNI), K-Nearest Neighbours Selected by Correlation Imputation (KNNSCI) and Correlation Weighted Regression Imputation (CWRI). These techniques aim to tackle some of the drawbacks found in the methods from literature, such as the necessity for a complete subset, evaluation on solely one missingness mechanism, and valid imputation only for numeric variables.

Furthermore, they exploit the concept of correlation, a promising but still not widely adopted approach in medical research. Rather than just using the correlation between values, these methods investigate the potential benefits of considering the correlation between values and missingness patterns, an innovative and unique strategy. The following three sections outline the framework of CWKNNI, KNNSCI and CWRI. The correlation coefficients selected after the first stage of the correlational study, described in Section 5.3, are used throughout this section. Note that, since the core concept of these three methods is similar, the first steps of their implementation are identical because they mostly concern the calculation of correlation matrices.

5.5.2.1 Correlation Weighted K-Nearest Neighbour Imputation

The CWKNNI method was inspired by the CoHiKNN algorithm, a KNN approach that utilises the correlation between attributes as weights in the calculation of the distances from each complete record to the target incomplete record. However, instead of uniquely considering the correlation between values, the weights are obtained through a weighted average of that correlation and the correlation between values and missingness pattern. Additionally, CWKNNI overcomes the limitation of CoHiKNN in terms of its dependency

on a complete data subset to impute the missing values, as it computes the correlation matrix through a pairwise deletion approach instead of performing listwise deletion.

A simple flowchart of CWKNNI is shown in Figure 5.2(a). Furthermore, the following step-by-step explanation provides the outline for this method:

1. Consider a dataset X, with f features and N instances. Additionally, let f_{miss} denote the number of attributes with missing values.

2. Compute the $f \times f$ correlation matrix C_{vv} adopting a pairwise deletion strategy, i.e. calculate the correlation between the available values within each pair of attributes on an analysis-by-analysis basis. In addition, calculate a $f_{\text{miss}} \times f$ matrix, hereby denoted as C_{vm}, with the correlations between values and missingness patterns, according to the procedure described in Section 5.3.2. In this approach, the absolute value of these correlations was considered, thus accounting for the strength of the association, not its direction.

3. Order the features from the one with the lowest MR to the one with the highest. Imputation will be performed in this sequence, in a phased manner.

4. Let f_j denote the attribute being imputed, where j is its index. Create a subset $X_{\text{miss\_}j}$ consisting of samples where the attribute f_j is missing. Furthermore, create a pool of samples where the attribute f_j is observed.

5. For each instance in $X_{\text{miss\_}j}$ find the k nearest neighbours within the subset's respective pool. A weighted euclidean distance which accounts for the presence of missing values is used as a distance measure:

$$d_{vt} = \sqrt{w_{\text{D}} \times \sum_{i \in O_f} (1 - w_{\text{C}i}) \times (v_i - t_i)^2} \tag{5.1}$$

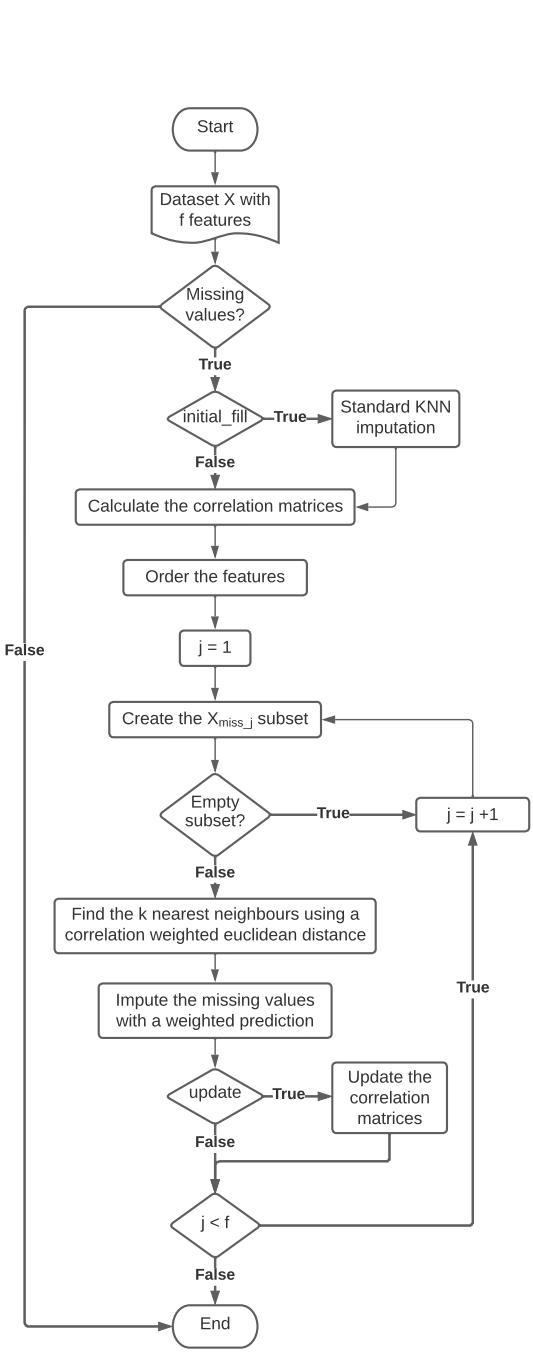
where d_{vt} is the distance between samples v and t, v is an instance of $X_{\text{miss\_}j}$, t is a sample from the corresponding pool, v_i and t_i are the observed values from the ith attribute of v and t respectively, and O_f is the set of indexes of the variables that are not missing neither in v nor in t. w_{D} is the quotient between the total number of features f and the dimension of O_f, i.e. the number of features observed in both v and t. As for $w_{\text{C}i}$, it is a weighted average of the correlation between f_j and f_i, and the correlation between the values of f_i and the missingness pattern of f_j:

$$w_{\text{C}i} = p \times C_{\text{vv}}[i, j] + (1 - p) \times C_{\text{vm}}[i, j] \,, \qquad p \in [0, 1] \tag{5.2}$$

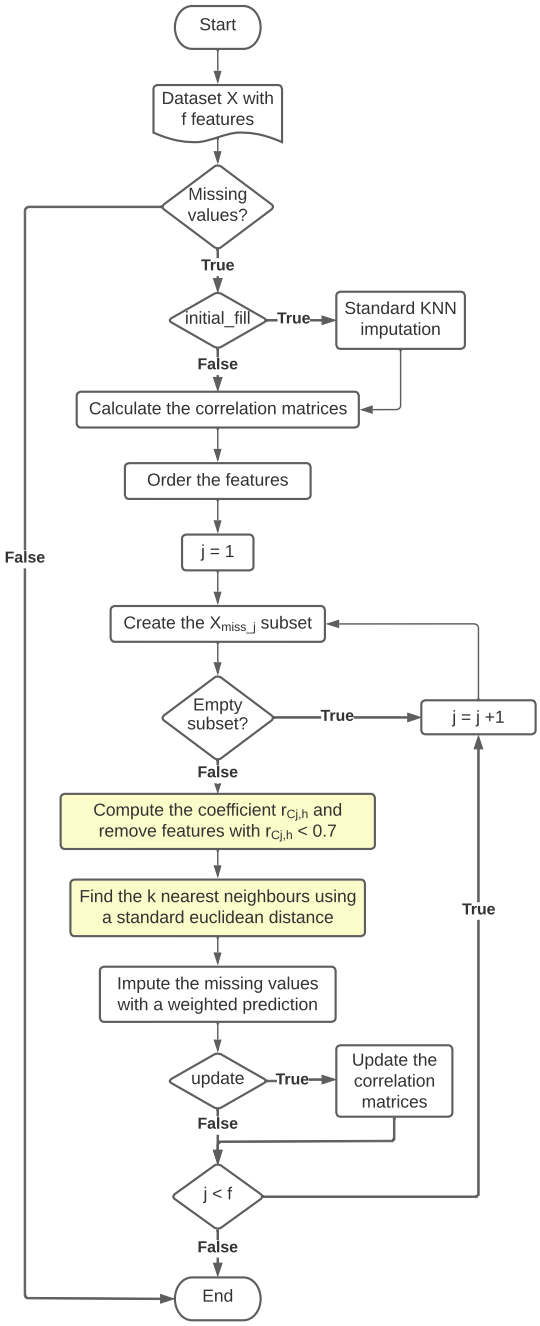
Note that $0 \leq w_{\text{C}i} \leq 1$. The term $(1 - w_{\text{C}i})$ in Equation 5.1 ensures that a stronger $w_{\text{C}i}$ leads to a the greater fading of $(v_i - t_i)^2$, lowering the computed distance.

6. Replace the missing value in each instance of $X_{\text{miss\_}j}$ with a weighted prediction, in which the weights are the inverse of the computed distances d_{vt}: a closer neighbour has a higher importance (i.e. weight) in the final prediction. The mean of the values is used for numeric variables, whereas the mode is used to impute binary attributes.

7. Repeat Steps 4-6 until all missing values, from all features, have been imputed.



(a) Flowchart of CWKNNI



(b) Flowchart of KNNSCI - The differences in comparison to CWKNNI were highlighted in yellow

Figure 5.2: Flowcharts of the (a) CWKNNI and (b) KNNSCI methods.

The number of neighbours k and the percentage p, i.e. the weight placed on the correlation between values in comparison to the correlation between values and missingness pattern, are two parameters of CWKNNI.

This method also accepts a boolean parameter, **initial_fill**, which determines if an initial and temporary imputation with the standard KNN method is carried out. The values from this imputed dataset will be used for the calculation of both C_{vv} and C_{vm}. Furthermore, the pool for each subset X_{miss_j} will be formed by these KNN imputed samples. The number of samples within each pool remains the same, the only difference is that no instance has missing attributes. As for the sample being imputed, f_j will be its only missing feature.

Additionally, the boolean argument **update** controls if the imputed values will be used in subsequent phases of imputation, instead of just considering the original values of each sample. Both matrices C_{vv} and C_{vm} are updated at the beginning of each phase, thus accounting for the newly imputed values. Note that the missingness pattern used to calculate C_{vm} does not change.

In conclusion, CWKNNI presents a total of four parameters, and the values tested for each one are shown in Table 5.3.

Table 5.3: Parameter values for the proposed approaches.

Method	Parameters
CWKNNI	**n_neighbors** $\in \{5, 10, 15\}$, **percentage** $\in \{0.2, 0.4, 0.6, 0.8\}$, **initial_fill** $\in$ {False, True}, **update** $\in$ {False, True}
KNNSCI	**n_neighbors** $\in \{5, 10, 15\}$, **percentage** $\in \{0.2, 0.4, 0.6, 0.8\}$, **initial_fill** $\in$ {False, True}, **update** $\in$ {False, True}
CWRI	**percentage** $\in \{0.2, 0.4, 0.6, 0.8\}$

5.5.2.2 K-Nearest Neighbours Selected by Correlation Imputation

As most KNN-based approaches, CWKNNI exhibits an increased computational cost for high dimensional data. To overcome this drawback, the KNNSCI method was created. This technique performs a pre-selection of features based on correlation, which reduces the dimensionality and facilitates its application on large datasets. As mentioned before, KNNSCI accounts for both the correlation between values, and the correlation between values and missingness pattern.

Figure 5.2(b) provides a simple flowchart of KNNSCI. Morevoer, the outline for this method is given in the step-by-step description below:

1. Consider a dataset X, with f features and N instances. Additionally, let f_{miss} denote the number of attributes with missing values.

2. Compute the $f \times f$ correlation matrix C_{vv} adopting a pairwise deletion strategy, i.e. calculate the correlation between the available values within each pair of attributes on an analysis-by-analysis basis. In addition, calculate a $f_{miss} \times f$ matrix, hereby denoted as C_{vm}, with the correlations between values and missingness patterns, according to the procedure described in Section 5.3.2. In this approach, the absolute value of these correlations was considered, thus accounting for the strength of the association, not its direction.

3. Order the features from the one with the lowest MR to the one with the highest. Imputation will be performed in this sequence, in a phased manner.

4. Let f_j denote the attribute being imputed, where j is its index. Create a subset X_{miss_j} consisting of samples where the attribute f_j is missing.

5. Using the correlation matrices, compute the coefficient $r_{Cj,h}$ between attribute f_j and the hth attribute of X (apart from f_j). This coefficient is equal for all samples within the subset X_{miss_j} and is obtained through the following equation:

$$r_{Cj,h} = p \times C_{vv}[h, j] + (1 - p) \times C_{vm}[h, j] \,, \quad p \in [0, 1] \tag{5.3}$$

   Note that $0 \leq r_{Cj,h} \leq 1$.

6. Create a pool of samples where the attribute f_j is observed. Find the features where $r_{Cj,h} < 0.7$ and remove their columns from both X_{miss_j} and corresponding pool. The value of 0.7 was chosen based on the work of Schober, Boer, and Schwarte [17], which stated that a coefficient above 0.7 indicates a strong correlation.

7. For each instance in X_{miss_j} find the k nearest neighbours within the subset's respective pool (after removing the columns with $r_{Cj,h} < 0.7$). A standard euclidean distance is used as a distance measure.

8. Replace the missing value in each instance of X_{miss_j} with a weighted prediction, in which the weights are the inverse of the computed euclidean distances: a closer neighbour has a higher importance in the final prediction. The mean of the values is used for numeric variables, whereas the mode is used to impute binary attributes.

9. Repeat Steps 4-8 until all missing values, from all features, have been imputed.

In addition to the number of neighbours k and the percentage p, KNNSCI also has the parameters **initial_fill** and **update**, which serve a similar purpose as the parameters with the same name in CWKNNI. Table 5.3 exhibits the tested values for each parameter of the KNNSCI algorithm.

5.5.2.3 Correlation Weighted Regression Imputation

The CWRI approach is a regression-based method, in which the missing values are imputed with predictions drawn from distinct linear regression models. This technique finds several estimates for each missing value, and combines them taking into account the correlational importance of the predictor variables. This correlational importance is obtained through a weighted average of the correlation between values and the correlation between values and missingness pattern.

Figure 5.3 displays a simple flowchart of CWRI. Similar to the previous sections, a step-by-step outline of CWRI is presented:

1. Consider a dataset X, with f features and N instances. Additionally, let f_{miss} denote the number of attributes with missing values.

2. Compute the $f \times f$ correlation matrix C_{vv} adopting a pairwise deletion strategy, i.e. calculate the correlation between the available values within each pair of attributes on an analysis-by-analysis basis. In addition, calculate a $f_{\text{miss}} \times f$ matrix, hereby denoted as C_{vm}, with the correlations between values and missingness patterns, according to the procedure described in Section 5.3.2. In this approach, the absolute value of these correlations was considered, thus accounting for the strength of the association, not its direction.

3. Build a $f \times f$ matrix, denoted as R, containing multiple linear regression models. Let f_i and f_j be two different attributes of X, with indexes i and j respectively. For every pair $\{i, j\}$ (where $i \in \{1, 2, \ldots, f\}$, $j \in \{1, 2, \ldots, f\}$, and $i \neq j$), $R[i, j]$ stores a linear regression model in which f_i is the independent variable and f_j is the predictor.

4. Assemble the subset X_{incomp} including all incomplete samples, i.e. instances with at least one missing value.

5. Consider any instance of X_{incomp}, and let f_k denote its kth missing attribute. Using every non-missing feature f_t in this sample, obtain the corresponding predicted value for f_k through the regression model $R[k, t]$. If f_k is a numeric variable, the final imputation will be a weighted average of all predicted values. If f_k is binary, a weighted mode is applied. To obtain the weight given to each variable f_t, denoted as $w_{k,t}$, first calculate its correlational importance $w_{Ck,t}$ through the following equation:

$$w_{Ck,t} = p \times C_{\text{vv}}[t, k] + (1 - p) \times C_{\text{vm}}[t, k] \,, \quad p \in [0, 1] \tag{5.4}$$

   Note that $0 \leq w_{Ck,t} \leq 1$. After computing $w_{Ck,t}$ for every non-missing feature, $w_{k,t}$ is a simple normalisation:

$$w_{k,t} = \frac{1}{\sum_t w_{Ck,t}} \times w_{Ck,t} \tag{5.5}$$

6. Repeat Step 5 until all incomplete samples have been imputed.

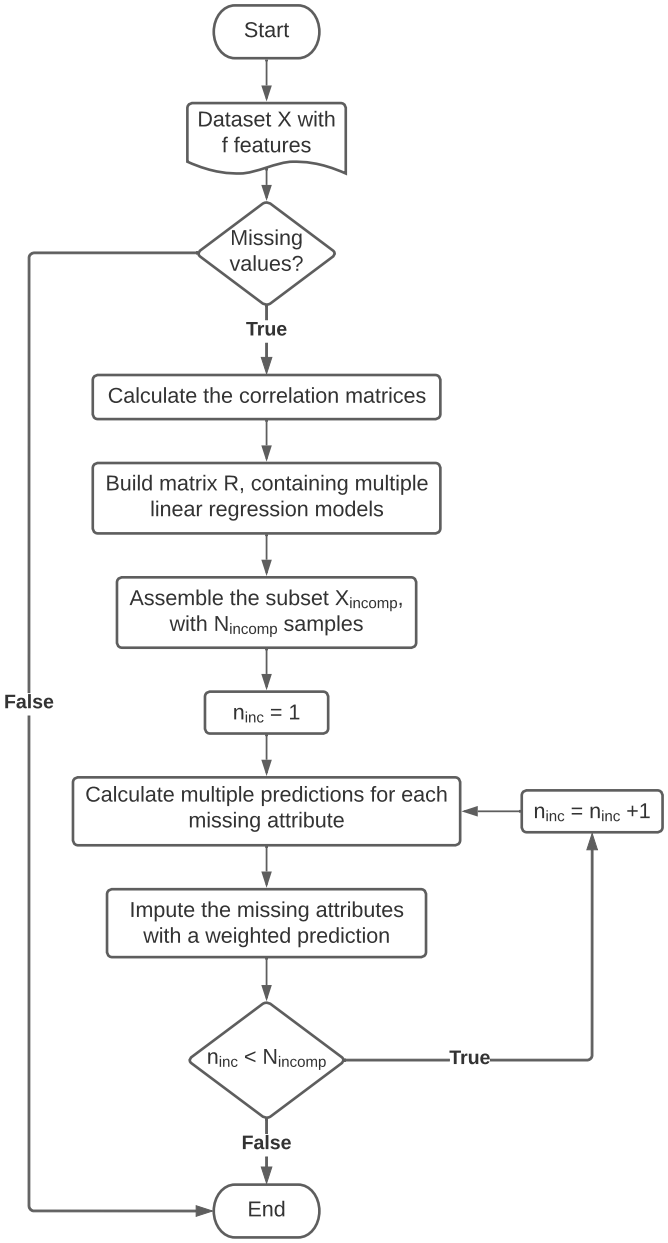


Figure 5.3: Flowchart of the CWRI method.

The percentage p, i.e. the weight placed on the correlation between values in comparison to the correlation between values and missingness pattern, is the only parameter of CWRI. The tested values for this parameter are presented in Table 5.3.

5.6 Training Pipeline

After performing missing value imputation, three ML models were trained: a RF classifier, a SVM classifier, and a NB classifier. The procedure followed for the datasets with injected missing values is detailed first, and the procedure for the real-world medical datasets

after, since the adopted strategy differed. Both procedures used a grid search with 5-fold cross validation for the optimisation of the models through hyperparameter tuning.

As illustrated in Figure 5.1(a), a grid search technique is applied to each complete UCI Machine Learning Repository databases, instead of the imputed datasets. This choice greatly reduces the computational cost, since the alternative would be to run a grid search for each stratified fold of all imputed dataset. Hence, for every UCI Machine Learning Repository dataset, one grid search algorithm is computed using the 5 stratified folds for the 5-fold cross validation strategy. The tested hyperparameter values for each model are detailed in Table 5.4. For each one of the three ML classifiers, the optimal hyperparameters are the ones from the model with the highest average AUROC score. This metric was chosen because it is widely-used in medical research to evaluate clinical prediction models, as a greater AUROC indicates that the model classifies more instances correctly.

Table 5.4: Tested hyperparameter values for the selected ML classifiers in the framework of the synthetically generated datasets.

Classifier	Hyperparameters
RF	**random_state** = 42, **max_depth** $\in \{3, 5, 7, 9\}$, **n_estimators** $\in \{5, 10, 25, 50, 100, 200\}$, **criterion** $\in$ {'entropy', 'gini'}, **min_samples_split** $\in \{10, 20, 50\}$, **min_samples_leaf** $\in \{5, 10, 25\}$
SVM	**random_state** = 42, **class_weight** = 'balanced', **C** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$, **gamma** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$
NB	**var_smoothing** $\in \{1.0 \times 10^{-4}, 2.8 \times 10^{-5}, 7.7 \times 10^{-6}, 2.2 \times 10^{-6}, 6.0 \times 10^{-7}, 1.7 \times 10^{-7}, 4.6 \times 10^{-8}, 1.3 \times 10^{-8}, 3.6 \times 10^{-9}, 1.0 \times 10^{-9}\}$

The procedure followed for the real-world datasets was slightly different, as depicted in Figure 5.1(b). In this case, as there is no baseline dataset (ground truth), it was considered necessary to compute a grid search with 5-fold cross validation for each stratified fold of each imputed dataset to ensure a fair comparison between the imputation techniques. In order to soften the computational cost, the number of tested values within some hyperparameters of the RF classifier was lowered, as shown in Table 5.5. Similarly to the above case, the optimal model in each computed grid search was the one that achieved the highest AUROC.

The performed comparative study aimed to be as rigorous and fair as possible. To this end, hyperparameter tuning constituted an essential step to guarantee that each imputation technique was being evaluated under its best conditions, minimising the likelihood of any factors external to the imputation procedure corrupting the results.

Table 5.5: Tested hyperparameter values for the selected ML classifiers in the framework of the real-world medical datasets.

Classifier	Hyperparameters
RF	**random_state** = 42, **max_depth** $\in \{3, 6, 9\}$, **n_estimators** $\in \{10, 25, 100\}$, **criterion** $\in$ {'entropy', 'gini'}, **min_samples_split** $\in \{20, 50\}$, **min_samples_leaf** $\in \{10, 25\}$
SVM	**random_state** = 42, **class_weight** = 'balanced', **C** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$, **gamma** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$
NB	**var_smoothing** $\in \{1.0 \times 10^{-4}, 2.8 \times 10^{-5}, 7.7 \times 10^{-6}, 2.2 \times 10^{-6}, 6.0 \times 10^{-7}, 1.7 \times 10^{-7}, 4.6 \times 10^{-8}, 1.3 \times 10^{-8}, 3.6 \times 10^{-9}, 1.0 \times 10^{-9}\}$

5.7 Performance Evaluation

As aforementioned, two types of performance evaluation were carried out: imputation evaluation and classification evaluation.

The first type assessed the quality of the imputation procedure by comparing the estimated values with the original ones. This evaluation could only be implemented on the incomplete datasets generated from the three UCI Machine Learning Repository datasets, since it is impossible to trace back the real values of the missing elements in the remaining datasets. Section 2.4.2.4 presented different criteria to assess the precision of the imputation procedure, from which the MAE was chosen for the evaluation.

The classification evaluation studied the impact of each imputation procedure on a ML model's performance, namely on a RF classifier, a SVM classifier, and a NB classifier. In addition, these models were also trained upon datasets to which listwise deletion was applied. A 5-fold cross validation was carried out using the stratified split mentioned in Section 5.4. The AUROC was selected to assess each classifier's performance due to its widespread use in the evaluation of clinical prediction models.

Chapter 6

Results and Discussion

This chapter presents and discusses the various results obtained within this dissertation, following the methodologies previously outlined in Chapter 5. The comprehensive studies that were carried out aimed to answer the RQs posed in Chapter 1 regarding correlation and missing value imputation.

6.1 Correlational Study

In the first stage of this correlational study, a comparison between correlation measurements was performed. This stage determined which correlation coefficients would be used in further analyses. The second stage explored the relationships between variables and missingness patterns of each missingness mechanism.

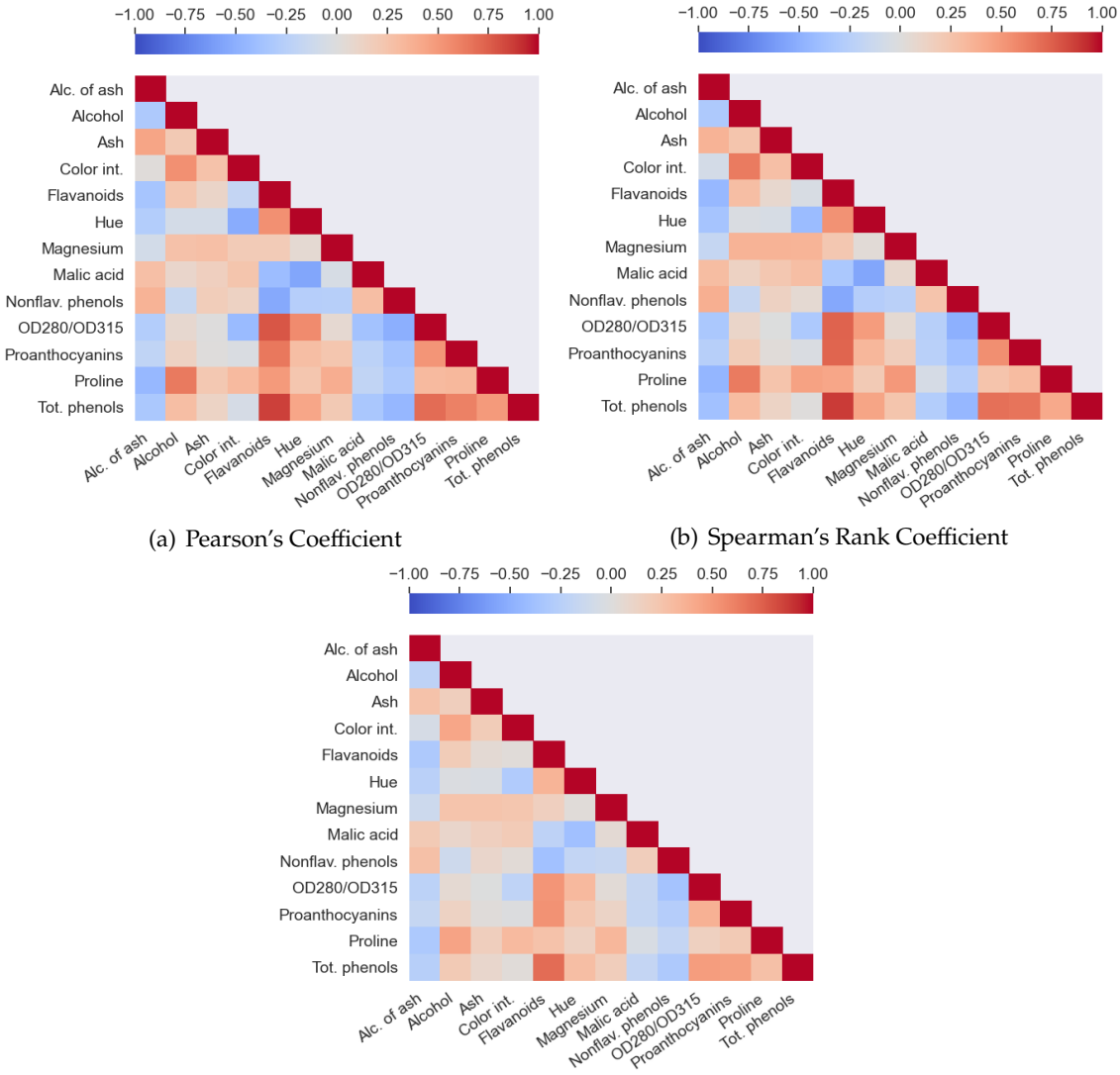
6.1.1 Correlation Coefficients Comparison

Taking into account the types of variables considered throughout this dissertation, it was necessary to select three coefficients, one for each type of relationship (numeric-numeric, binary-binary, and binary-numeric). Moreover, the sole objective of this stage is to compare the values provided by different correlation coefficients, which does not necessarily entail their interpretation.

As for coefficients that measure the correlation among two numeric attributes, three measurements were compared: Pearson's coefficient, Spearman's rank coefficient, and Kendall's Tau coefficient. These three statistical measurements range from -1 to 1 but have distinct underlying assumptions, as discussed in Section 2.2.1.2. Pairwise correlation was computed for all numeric variables of the Wine Data Set and the Statlog (Heart) Data Set. Since these coefficients measure symmetric and mutual associations, the correlation between any two variables X and Y is equal to the one between Y and X, i.e. the two

variables are interchangeable. Therefore it is only necessary to calculate the correlation for one of these and any two such pairings.

Figure 6.1 depicts the lower-triangle of the correlation matrices obtained for the Wine Data Set through these three coefficients. The upper-triangle of the matrices was omitted for legibility purposes, as the information provided is redundant. In terms of pattern, the Pearson's and Spearman's rank coefficients are the closest, with Kendall's Tau coefficient generally showing lower absolute values, as discussed in Section 2.2.1.2. The largest discrepancy between the modulus of the Pearson's and Spearman's coefficients is 0.16, found when measuring the correlation among the attributes "Magnesium" and "Color int.". Similar inferences are drawn from the correlation matrices computed for the numeric attributes of the Statlog (Heart) Data Set, depicted in Figure B.1. In this case, the largest difference between the Pearson's and Spearman's coefficients is 0.07, an even lower value.



(a) Pearson's Coefficient

(b) Spearman's Rank Coefficient

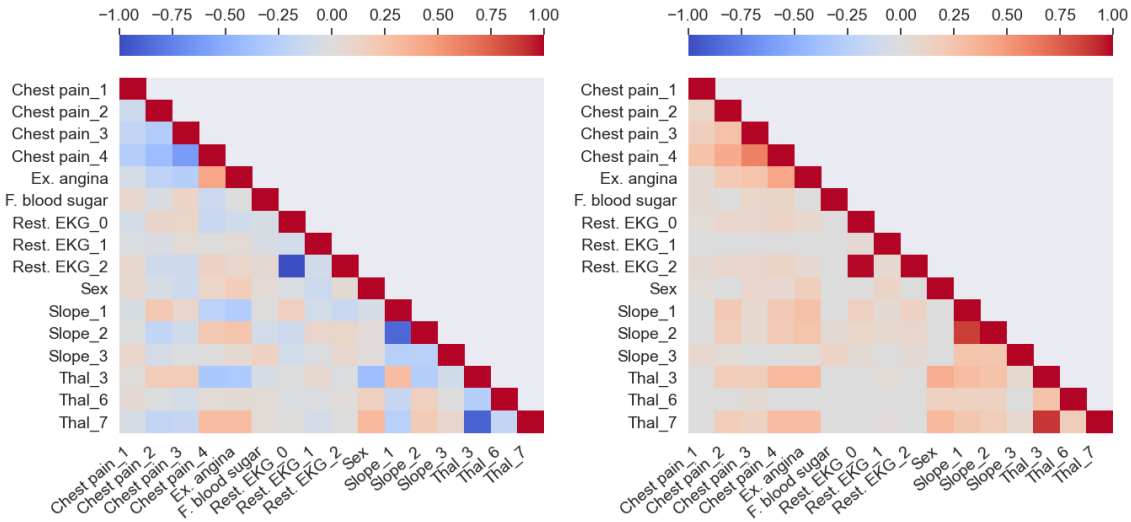
(c) Kendall's Tau Coefficient (B)

Figure 6.1: Correlation matrices of the Wine Data Set obtain through different coefficients: (a) Pearson's coefficient, (b) Spearman's Rank coefficient, (c) Kendall's Tau coefficient.

When compared to Kendall's Tau, Spearman's rank coefficient yields more reliable inferences in the presence of tied ranks, as mentioned in Section 2.2.1.2. Real-world medical datasets are usually high dimensional and prone to tied ranks, and thus Kendall's Tau was not chosen. Given that there are no significant disparities between the Pearson's and the Spearman's coefficient, the former was selected as it is the most used in research.

Moreover, regarding the correlation among binary variables, two coefficients were compared: Phi coefficient, and Cramér's V (with its bias correction). As stated in Section 2.2.1.2, the uncorrected Cramér's V is an extension of the Phi coefficient, and therefore the correlations measured by these two coefficients are expected to be similar in modulus. The main difference between these metrics is the fact that Phi ranges from -1 to 1, while Cramér's V only takes positive values up to 1. Pairwise correlation was computed for all binary variables of the Statlog (Heart) Data Set and the SPECT Heart Data Set.

Figure 6.2 contains the lower-triangle of the two correlation matrices calculated for the binary attributes of the Statlog (Heart) Data Set through the Phi coefficient and the Cramér's V (corrected). In terms of pattern, the clear differences are due to the distinct ranges of these two coefficients, i.e. to the fact that Cramér's V is never negative. Although the matrix in Figure 6.2(b) does not exhibit negative values, it is possible to establish a relationship with the matrix in Figure 6.2(a) in terms of absolute value: the higher the negative correlation measured by the Phi coefficient, the higher the positive value of Cramér's V. If the modulus of the correlations given by the Phi Coefficient are compared to the ones computed by Cramér's V, the largest disparity takes the value of 0.06 on a scale of 0 to 1, which is a non-significant difference. A similar behaviour is observed for the SPECT Heart Data Set, whose matrices are shown in Figure B.2. In this case, the largest disparity takes the value of 0.07, which is non-significant on a scale of 0 to 1.



(a) Phi Coefficient

(b) Cramér's V (corrected)

Figure 6.2: Correlation matrices of the binary variables of the Statlog (Heart) Data Set obtained through different coefficients: (a) Phi Coefficient, (b) Cramér's V (corrected).

Since there are no considerable discrepancies between the correlations measured by the Phi coefficient and Cramér's V, the former was chosen for this dissertation because it is more widely-used for binary attributes and is able to assess the direction (positive or negative) of an association, whereas the latter only provides an absolute value.

Finally, the point biserial coefficient will be used to measure the correlation between a numeric and a binary attribute. Recall that this coefficient was automatically selected as there are no other suitable alternatives. To summarise, this initial study led to the selection of three distinct correlation coefficients to deal with different types of variables: Pearson's coefficient, Phi coefficient and point biserial coefficient. These metrics are all based on Pearson's correlation, and therefore the correlation values obtained for different variable types are comparable.

6.1.2 Correlation Between Values

The second stage of the correlational study initiated with an evaluation concerning the impact of both the missingness mechanism and MR on the correlation between the values of two variables. With that objective in mind, the correlation matrix from every dataset containing injected missing values was compared with the correlation matrix from its corresponding original (and complete) dataset. The MAE was used to measure the average difference between the correlations of every pair of corresponding elements from these matrices. Note that pairs in which none of the variables were missing were not included in the calculation of this error. The results are shown in Table 6.1.

Table 6.1: Impact of the missingness mechanism and MR on the correlation between values assessed by the MAE. For legibility purposes, the values are multiplied by 10^2.

		(MAE ± Standard Deviation) $\times 10^2$		
		10%	**30%**	**50%**
Wine Data Set	MCAR	3.38 ± 3.07	6.42 ± 4.92	9.17 ± 7.90
	MAR	3.45 ± 3.27	10.73 ± 11.22	17.49 ± 18.64
	MNAR	5.84 ± 3.82	20.81 ± 14.56	28.55 ± 21.03
SPECT Heart Data Set	MCAR	3.06 ± 2.20	6.79 ± 4.80	9.72 ± 7.26
	MAR	2.52 ± 1.85	5.75 ± 4.30	9.41 ± 8.18
	MNAR	3.39 ± 2.27	8.86 ± 7.56	14.60 ± 14.27[(a)]
Statlog (Heart) Data Set	MCAR	1.95 ± 1.74	4.59 ± 3.93[(a)]	7.93 ± 6.76[(a)]
	MAR	1.57 ± 1.46	4.45 ± 3.95[(a)]	7.05 ± 8.18[(a)]
	MNAR	2.15 ± 1.94	4.83 ± 4.68[(a)]	8.79 ± 9.58[(a)]

[(a)] This MAE does not include all possible pairs of variables because the missingness precluded the calculation of some correlation values.

In every dataset, for any missingness mechanism, the MAE grows as the MR increases. Furthermore, Figure 6.3 suggests that this relationship is approximately linear. A higher MR leads to less knowledge about the dataset, which increases the uncertainty when computing the correlation matrices, thus resulting in larger MAEs, i.e. a greater difference in comparison to the correlation values computed on the original dataset.

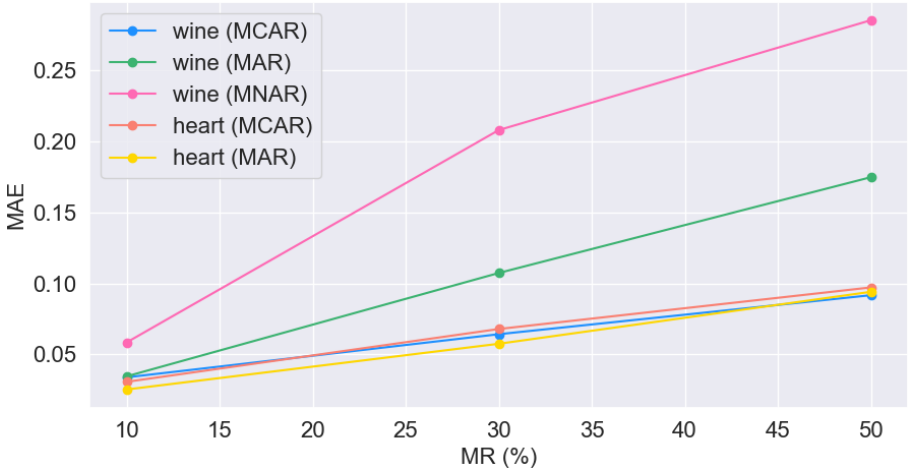
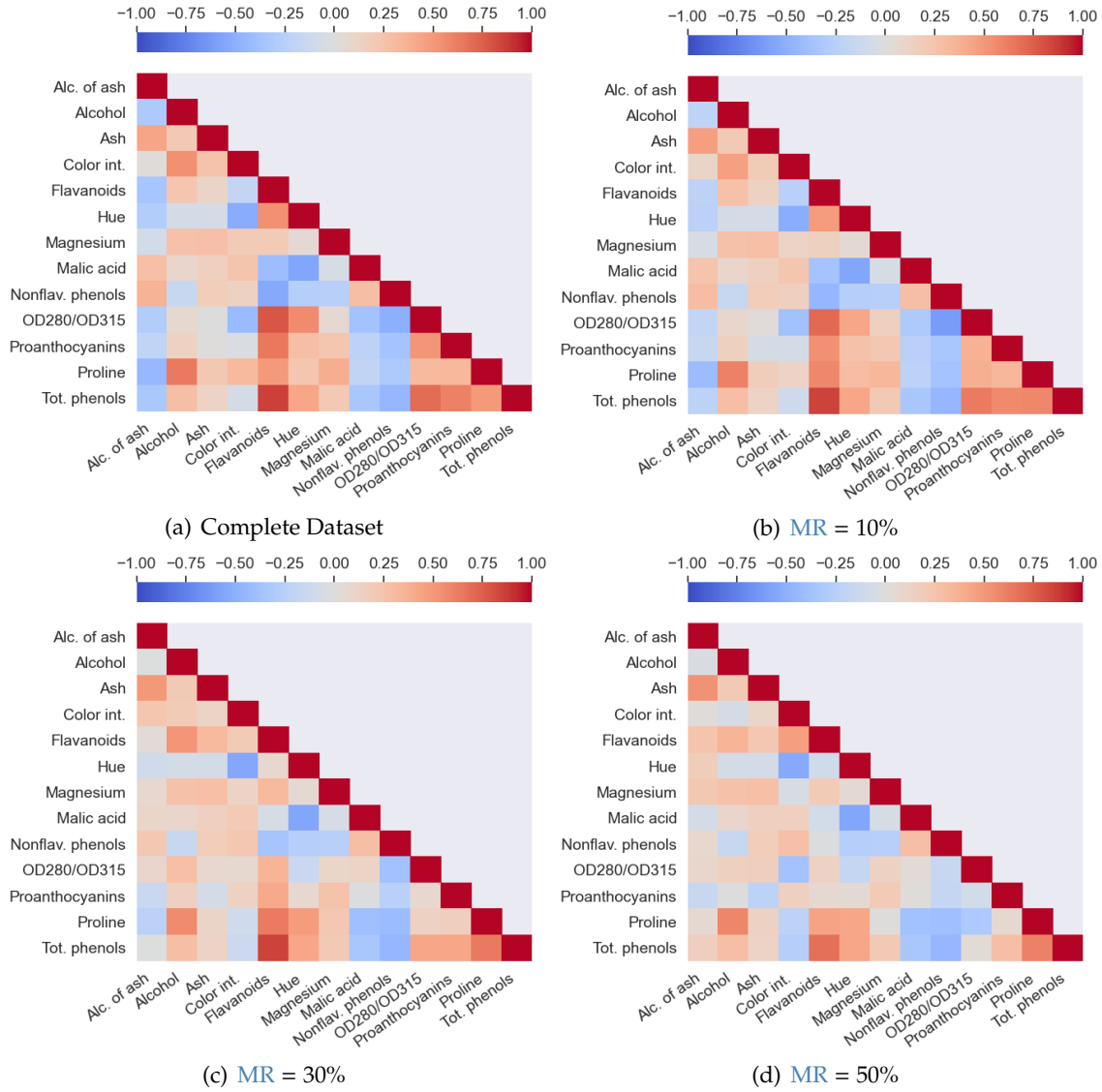


Figure 6.3: Graphical representation of Table 6.1. This graph only includes situations in which every MAE includes all possible pairs of variables.

As for the missingness mechanism, MNAR yields the highest MAEs on all three databases. Recall that, in this mechanism, the missing values are confined to the lower segment of their variable's distribution. The localised nature of this missingness has a greater impact on the correlations when compared to mechanisms in which the removal of elements is of a more haphazard nature. An entirely random missingness, i.e. MCAR, would not affect the computed correlations as much because the missing variable would still be representative of the originally (or hypothetically) complete variable.

However, the MCAR mechanism only generates the lowest errors on the Wine Data Set, although the MAEs are close to the ones from the MAR mechanism in the remaining two databases. Despite MCAR missingness being more haphazard than MAR, the generated MCAR datasets have a higher total missingness (view Section 4.1), which increases the uncertainty when calculating the correlation values. Consequently, the MAE of the MAR mechanism was slightly lower on the SPECT Heart Data Set and the Statlog (Heart) Data Set. These databases include binary attributes, not present in the Wine Data Set, which may indicate that MCAR and MAR missingness have similar effects in this type of variable.

Figure 6.4 contains the correlation matrices of the Wine Data Set obtained through the Pearson's coefficient for the original dataset and for different MRs under the MNAR mechanism. This visual representation makes it possible to notice that the correlation values soften as the MR grows, i.e. correlations that were strong in the complete dataset are weakened by the increase in the MR. Although the overall pattern maintains some similitude, high correlation values (in modulus) practically cease to exist for larger MRs. In fact, while the matrix in Figure 6.4(b) is nearly identical to the correlation matrix of the original dataset (Figure 6.4(a)), the matrix in Figure 6.4(d) no longer exhibits such significant similarities. A comparable behaviour is observed for the correlation matrices of the SPECT Heart Data Set under the MAR mechanism, shown in Figure B.3. The attenuation of the correlation values is not as prominent because the values regarding the complete dataset were not particularly strong.



(a) Complete Dataset

(b) MR = 10%

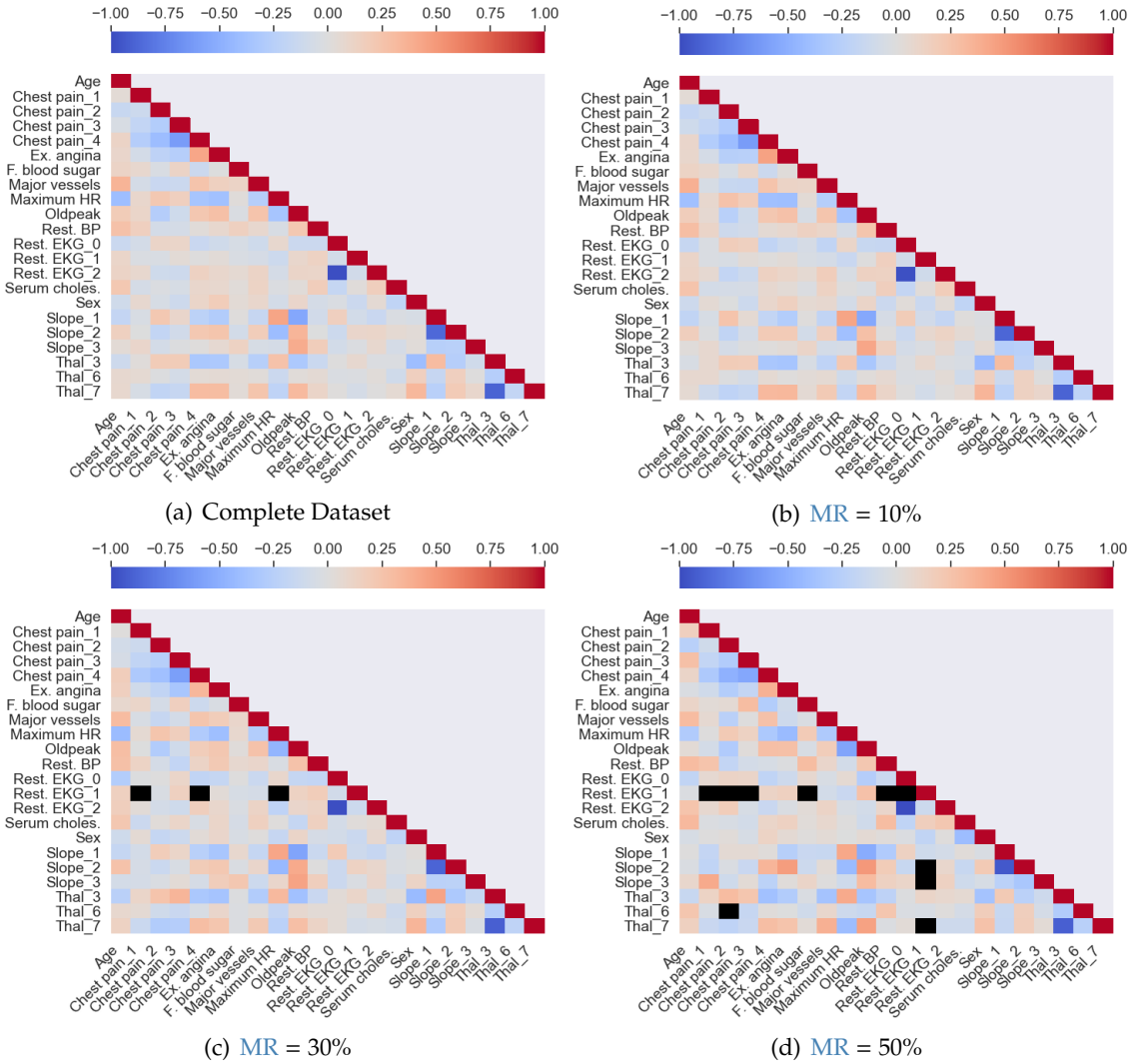
(c) MR = 30%

(d) MR = 50%

Figure 6.4: Correlation matrices of the Wine Data Set obtained for (a) the complete dataset and for different MRs under the MNAR mechanism: (b) MR = 10%, (c) MR = 30%, (d) MR = 50%.

On the other hand, the correlation matrices of the Statlog (Heart) Data Set under the MCAR mechanism, represented in Figure 6.5, do not exhibit a clear attenuation of the correlation values with the rise of the MR. As a matter of fact, Figure 6.5(d), concerning the highest MR, maintains the strong negative correlations between the variables "Rest. EKG_0" and "Rest. EKG_2", "Slope_1" and "Slope_2", and "Thal_3" and "Thal_7" for instance. Furthermore, the overall pattern remains nearly unchanged. An identical behaviour is observed in the Wine Data Set also under the MCAR mechanism, shown in Figure B.4. The disparity with the example above is due to the missingness mechanism, since the completely random nature of MCAR does not to affect computed correlations as greatly.

In summary, MNAR missingness has the most significant impact on the correlation



(a) Complete Dataset

(b) MR = 10%

(c) MR = 30%

(d) MR = 50%

Figure 6.5: Correlation matrices of the Statlog (Heart) Data Set obtained for (a) the complete dataset and for different MRs under the MCAR mechanism: (b) MR = 10%, (c) MR = 30%, (d) MR = 50%. The black squares correspond to situations in which the missingness precluded the calculation of the correlation.

between values, generating higher MAEs on all three databases. Although MAR missingness may yield lower MAEs in comparison with MCAR, the higher standard deviations indicate that the existing differences between correlation values are larger. In other words, the MAR mechanism generates fewer disparities between correlation matrices, but the ones that occur are greater than the ones from the MCAR mechanism.

Hence, the influence that missingness has on the correlation between values should not be overlooked when developing correlation-based imputation methods, particularly when working with datasets that have higher MRs. It may be important to consider a possible decrease in the correlation, even if this effect is only more prominent in MNAR, which some authors argue is the least prevalent mechanism in a real-world scenario [58].

6.1.3 Correlation Between Values and Missingness Pattern

This assessment aimed to investigate if correlation measures were capable of capturing the relationships between missingness patterns and observed values. For instance, the MAR elements synthetically inserted in one variable were determined by the observed values of another attribute. Therefore, the correlation between the missingness pattern of the first variable and the values of the second attribute should reflect a strong association. Although this is the most evident example, all three mechanisms were studied.

For every dataset with injected missing data, the matrix contemplating the correlations between values and missingness pattern, denoted as *var-miss* correlation matrix, was obtained. In order to investigate how these correlations vary between missingness mechanisms, Table 6.2 was analysed. Firstly, for each dataset, under each mechanism, the mean of all pairwise *var-miss* correlations from the three MRs was computed, resulting in a single *var-miss* correlation matrix with these averages. Then, the mean and standard deviation of the absolute value of all elements from every *var-miss* correlation matrix was calculated and is displayed in Table 6.2. Note that the standard deviations corresponding to the averages of all pairwise correlations from the three MRs are not considered.

Table 6.2: Average *var-miss* correlations obtained for three MRs under the same missingness mechanism. For legibility purposes, the values are multiplied by 10.

	(Mean ± Standard Deviation) ×10		
Dataset	**MCAR**	**MAR**	**MNAR**
Wine	0.47 ± 0.34	2.32 ± 1.74	2.24 ± 1.53
SPECT Heart	0.41 ± 0.32	1.23 ± 1.20	1.61 ± 1.32
Statlog (Heart)	0.35 ± 0.27	1.11 ± 1.24	1.09 ± 1.11

For every dataset, the MCAR mechanism yields the lowest average *var-miss* correlations, with no values greater than 0.05 on a scale of 0 to 1. This was expected given that MCAR missingness is entirely unrelated to the data, which leads to weaker *var-miss* correlations.

As for the MAR and MNAR mechanisms, their average *var-miss* correlations are practically identical and higher than MCAR's by one order of magnitude. Since MAR missingness is solely dependent on other measured variables, it is not surprising that the average *var-miss* correlations are more significant, ranging from 0.11 to 0.23 on a scale of 0 to 1. Although MNAR missingness does not necessarily have an association with the observed data, nothing precludes this from being present. Particularly, if the values of a missing attribute have a strong correlation with the values of another feature, then it is natural that the missingness of that attribute (which depends on its values) also has a strong correlation with the values of the second feature. The missing attributes in these three working datasets in fact have a significant correlation with at least one other variable, which explains the larger values of the computed *var-miss* correlations for MNAR.

Before delving into a visual inspection of the *var-miss* correlation matrices, it is important to first study if matrices under the same missingness mechanism differ greatly with

the MR. To this end, the standard deviation of all pairwise correlations across the three MRs was calculated for each dataset under each mechanism. Then, the average of the standard deviations of all pairs from the same dataset under a certain mechanism was computed and is shown in Table 6.3. The standard deviation was chosen because it is a measure of the dispersion around a mean value, and this assessment aims to analyse the variation of the *var-miss* correlations with the MR. Note that the MAEs could not be computed because there are no ground truth values for these correlations.

Table 6.3: Average of the standard deviations of the *var-miss* correlation matrices obtained for different MRs under the same missingness mechanism. For legibility purposes, the values are multiplied by 10^2.

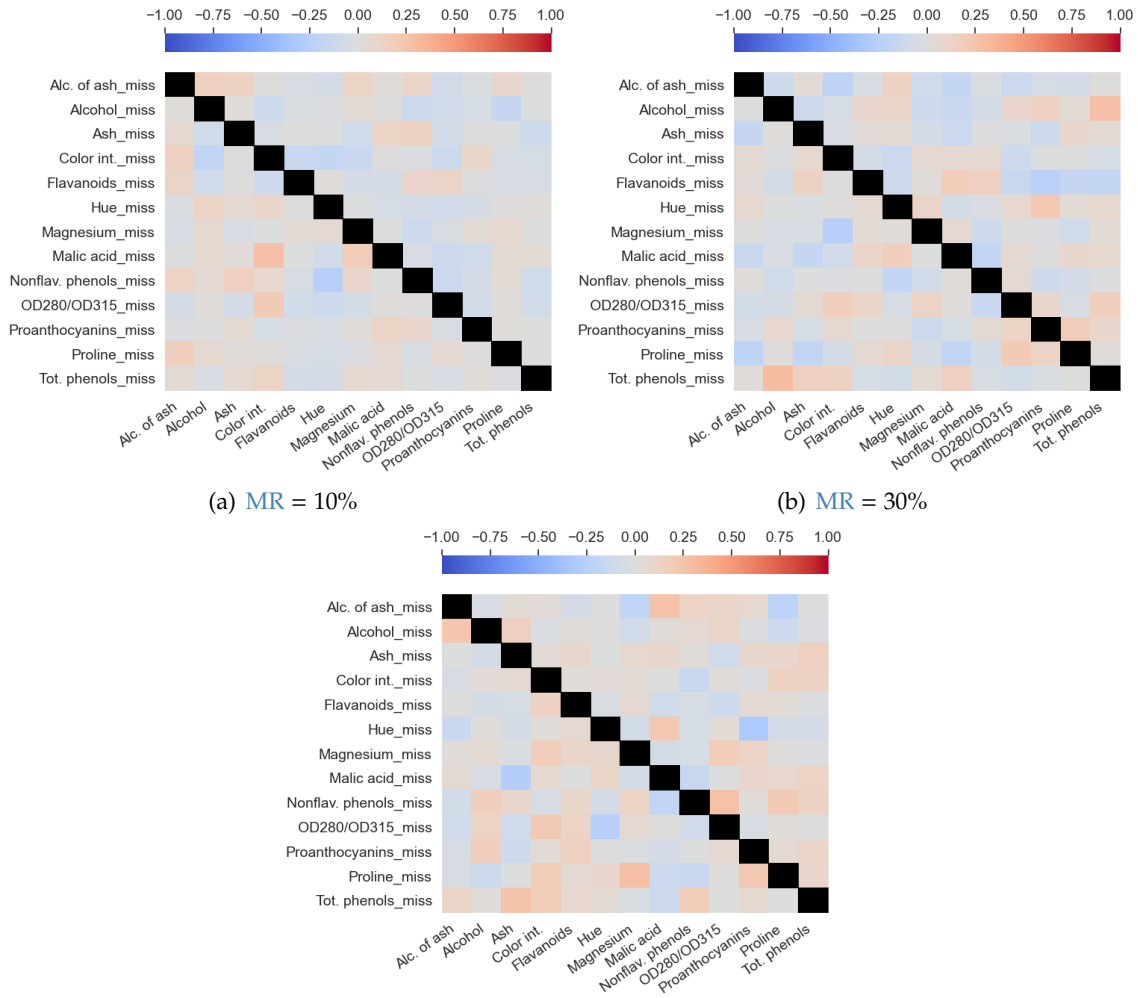
	(Mean ± Standard Deviation) $\times 10^2$		
Dataset	**MCAR**	**MAR**	**MNAR**
Wine	7.24 ± 3.82	7.83 ± 4.49	9.32 ± 4.89
SPECT Heart	6.48 ± 3.47	7.80 ± 4.81	7.90 ± 4.41
Statlog (Heart)	5.37 ± 2.89	5.31 ± 4.01	5.72 ± 4.00

In every dataset, the average of the standard deviations is higher for MNAR, followed by MAR and MCAR. However, the values are quite similar. Given that correlation ranges from -1 to 1 and that all values on Table 6.3 are below 0.10, no significant disparity is expected between matrices of different MRs under the same mechanism.

Next, some examples of *var-miss* correlation matrices from every missingness mechanism will be studied. Note that, contrary to the matrices concerning the correlation between values, the *var-miss* correlation matrices have different axes: the rows correspond to the missingness indicator for each missing variable and the columns to the values of each attribute in the dataset. For this reason, there is no redundant information in a matrix and it has to be completely represented (without removing the upper-triangle).

Figure 6.6 includes the *var-miss* correlation matrices for the Wine Data Set under the MCAR mechanism with different MRs. In this mechanism, the missingness is completely unrelated to the data, which explains why the computed correlations are overall weak for every MR, with no modulus value greater than 0.4. This also corroborates the low average *var-miss* correlations that the MCAR mechanisms presented in Table 6.2. Moreover, there does not seem to be a pattern that is consistent across MRs because the missing values were removed haphazardly. Whereas in the other mechanisms there is a visual trend that becomes more pronounced as the MR grows, MCAR does not show such regularity. This, along with the minimal variations in the *var-miss* correlations across the three MRs, might be causing the low standard deviations shown in Table 6.3. A similar behaviour is observed for the *var-miss* correlation matrices of the Statlog (Heart) Data Set under the MCAR mechanism, exhibited in Figure B.5. In this case, there is not a single *var-miss* correlation with an absolute value superior to 0.3.

As for the MAR mechanism, Figure 6.7 contains the *var-miss* correlation matrices for the SPECT Heart Data Set with different MRs. Contrary to the previous case, the



(a) MR = 10%

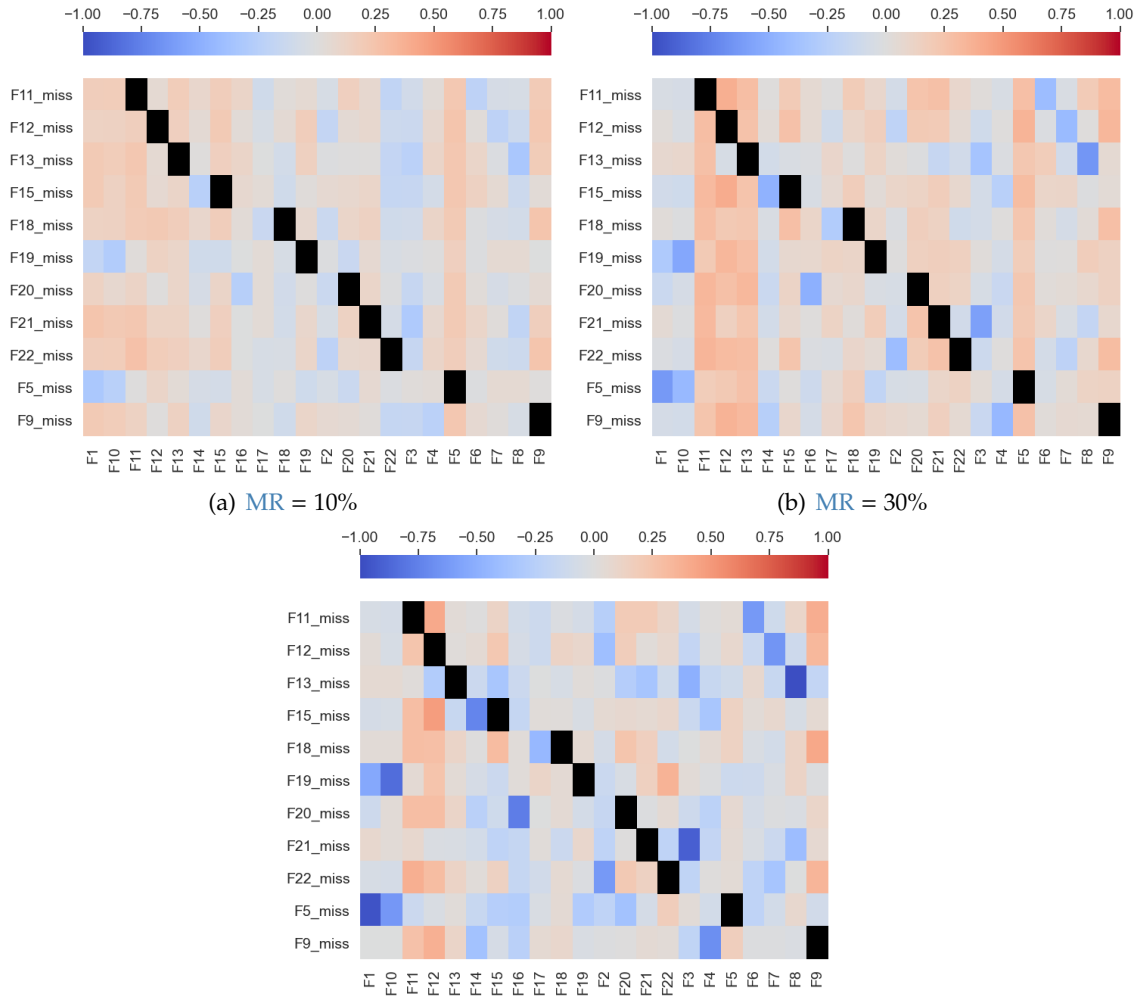
(b) MR = 30%

(c) MR = 50%

Figure 6.6: Correlation matrices between the values and the missingness pattern of the Wine Data Set obtained for different MRs under the MCAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%. Missing features have the suffix "_miss". The black squares correspond to correlations that are impossible to compute.

calculated correlations are fairly strong, particularly for higher MRs, which explains the greater *var-miss* correlations displayed in Table 6.2. There is an underlying pattern that becomes more pronounced as the MR increases. A larger number of missing values makes any relationship between observed data and missingness more evident, thus emphasising this pattern. Note that there are only a few pairs whose correlation has risen significantly in modulus across MRs, so it is natural that the average standard deviation shown in Table 6.3 is not high. For each missing attribute, i.e. each row of the matrix, there is always a negative value that stands out, corresponding to the correlation with the variable that determined the missingness of that attribute. The pair is negatively correlated because the missing elements on the non-determining feature coincide with the positions of the lowest values of the determining feature. Additionally, a missing feature sometimes has medium strength correlations with another attribute, for example “F19_miss” and “F1”.

This may indicate that there is a strong positive correlation between the values of "F1" and "F10", i.e. the determining feature of "F19_miss". In fact, this significant association is present in Figure B.3(a), and further inspection revealed it has a positive value of 0.66. The *var-miss* correlation matrices for the Wine Data Set, depicted in Figure B.6, also exhibit this behaviour under the MAR mechanism.



(a) MR = 10%

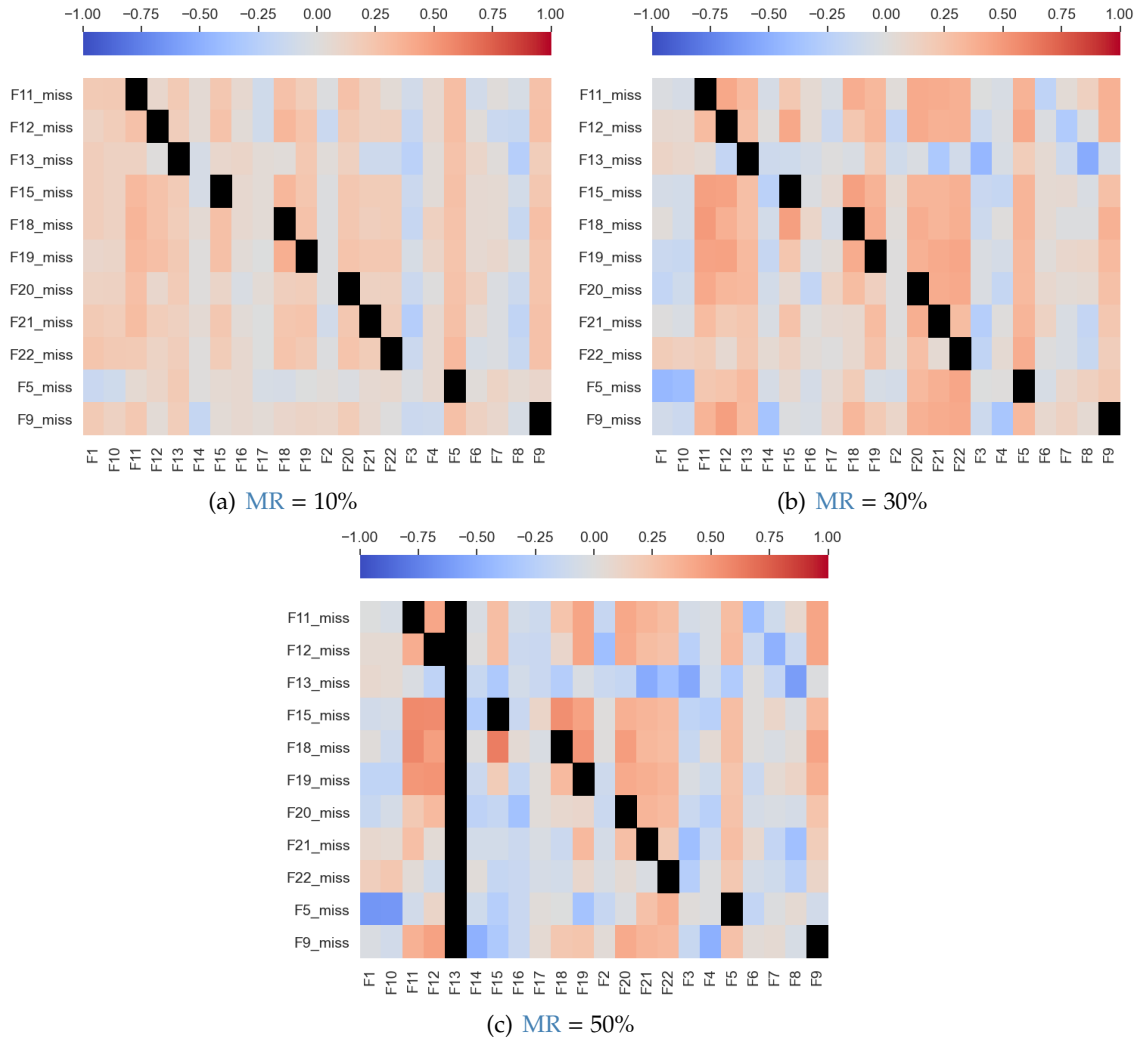
(b) MR = 30%

(c) MR = 50%

Figure 6.7: Correlation matrices between the values and the missingness pattern of the SPECT Heart Data Set obtained for different MRs under the MAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%. Missing features have the suffix "_miss". The black squares correspond to correlations that are impossible to compute.

Lastly, Figure 6.8 presents the *var-miss* correlation matrices for the SPECT Heart Data Set under the MNAR mechanism with different MRs. As in the situation above, the computed correlations have a reasonable strength even for lower MRs, coherent with the values of Table 6.2, and show an underlying pattern that is accentuated by a rise in the MR. Furthermore, there is not a sole correlation that stands out for each missing attribute, but rather a group composed of some correlations with a lower absolute value. For instance, "F13_miss" has a moderate correlation with "F21", "F3", and "F8", which may cause this

mechanism to be misinterpreted as MAR. In fact, MNAR missingness is related to the unobserved data of the missing variable alone. Hence, these moderate correlations are due to a spurious association between the missingness of "F13_miss" and the values of "F21", "F3", and "F8", caused by a significant correlation between the observed values of "F13" and the values of the other three variables, as exhibited in Figure B.3(a). This example proves that, in practice, it is impossible to confirm if the missingness in a certain variable is related to the unobserved data (MNAR) or is just a function of other measured variables (MAR). Similar inferences can be drawn from the *var-miss* correlation matrices of the Wine Data Set under the MNAR mechanism, included in Figure B.7.



(a) MR = 10%

(b) MR = 30%

(c) MR = 50%

Figure 6.8: Correlation matrices between the values and the missingness pattern of the SPECT Heart Data Set obtained for different MRs under the MNAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%. Missing features have the suffix "_miss". The black squares correspond to correlations that are impossible to compute.

In summary, the MCAR mechanism yields weak *var-miss* correlations that reflect the haphazard nature of this missingness, as expected. Furthermore, the *var-miss* correlation matrices do not show an underlying pattern across MRs as the missingness is unrelated to the measured data.

On the contrary, when computed for the MAR or MNAR mechanisms, these matrices display a consistent pattern across MRs, which includes strong correlations. These significant correlations were expected in the MAR mechanism, where one observed attribute determines the missingness in another. However, the presence of such correlations in the MNAR case are due to spurious associations and can lead to incorrect inferences concerning the identification of the mechanism.

Moreover, each MAR attribute in a *var-miss* correlation matrix, i.e. each row, exhibits a single stand-out correlation higher than the others, corresponding to the association with its determinant feature. In the MNAR case there is not a sole stand-out value in each row, but instead a group of moderate correlations. Alas, this disparity is not sufficiently strong to allow a distinction between the MAR and MNAR mechanisms.

Finally, note that the most noticeable discrepancies between the *var-miss* correlation matrices of all three mechanisms only become clear for higher MRs. Therefore, this assessment concluded that it is impossible to undoubtedly identify the missingness mechanisms solely using *var-miss* correlations. Albeit only particular cases have been discussed, the inferences drawn fairly generalise to any dataset under these three mechanisms.

6.1.4 Correlation Between Missingness Patterns

Lastly, the correlation between two missingness patterns, or nullity correlation, was studied in order to assess the degree to which one attribute's presence or absence influences another's presence. For example, when two variables are observed almost simultaneously, i.e. their missing values are in the same positions, they should have a strong positive nullity correlation. Furthermore, when the observation of one variable occurs concurrently with the absence of another, then their nullity correlation should be close to -1. These relationships were studied for all three missingness mechanisms.

The nullity correlation matrix was obtained for every synthetic dataset with injected missing data. As in the above section, it was first investigated whether the nullity correlations vary with the missingness mechanism. For each dataset, under each mechanism, the mean of all pairwise nullity correlations across the three MRs was calculated, resulting in a single matrix with these results. The average and standard deviation of the modulus of all elements from every generated matrix is compiled in Table 6.4. Note that the standard deviations of the means of all pairwise correlations from the three MRs are not considered.

MCAR has the lowest average nullity correlations in every dataset, with no absolute value higher than 0.04 on a scale of 0 to 1. These results are due to each MCAR variable exhibiting an haphazard missingness pattern, and therefore no strong association between these patterns is observed.

Table 6.4: Average nullity correlations obtained for three MRs under the same missingness mechanism. For legibility purposes, the values are multiplied by 10.

	(**Mean ± Standard Deviation**) ×10		
Dataset	**MCAR**	**MAR**	**MNAR**
Wine	0.33 ± 0.26	1.54 ± 1.02	2.25 ± 1.66
SPECT Heart	0.34 ± 0.24	3.80 ± 1.64	4.02 ± 1.56
Statlog (Heart)	0.28 ± 0.21	2.80 ± 2.34	2.27 ± 1.83

Regarding the MAR and MNAR mechanisms, their average nullity correlations are fairly similar and greater than the previous mechanism by one order of magnitude. In particular, MAR yields values from 0.15 in the Wine Data Set to 0.38 in the SPECT Heart Data Set, which indicates that moderate nullity correlations can occur. Recall that MAR missingness is not random by nature, as it relates to the values of a determining feature. Significant associations between missingness patterns of two distinct attributes may be observed if the values of the determining features of these attributes are correlated.

As for the MNAR mechanism, it shows values ranging from 0.23 in the Wine Data Set and Statlog (Heart) Data Set to 0.40 in the SPECT Heart Data Set. MNAR missingness is also not random by nature, depending on the values of the missing variable itself. Moderate associations may occur between missingness patterns of two distinct attributes if the values of the hypothetically complete attributes were correlated.

Afterwards, similar to the above section, the variability of the nullity correlation matrices across MRs was analysed. Thus, the standard deviation of all pairwise nullity correlations across the three MRs was obtained for each dataset and missingness mechanism. Table 6.5 contains the average of the standard deviations of all pairs from the same dataset under a certain mechanism. Once again, MAEs could not be computed because there are no ground truth values for the nullity correlations.

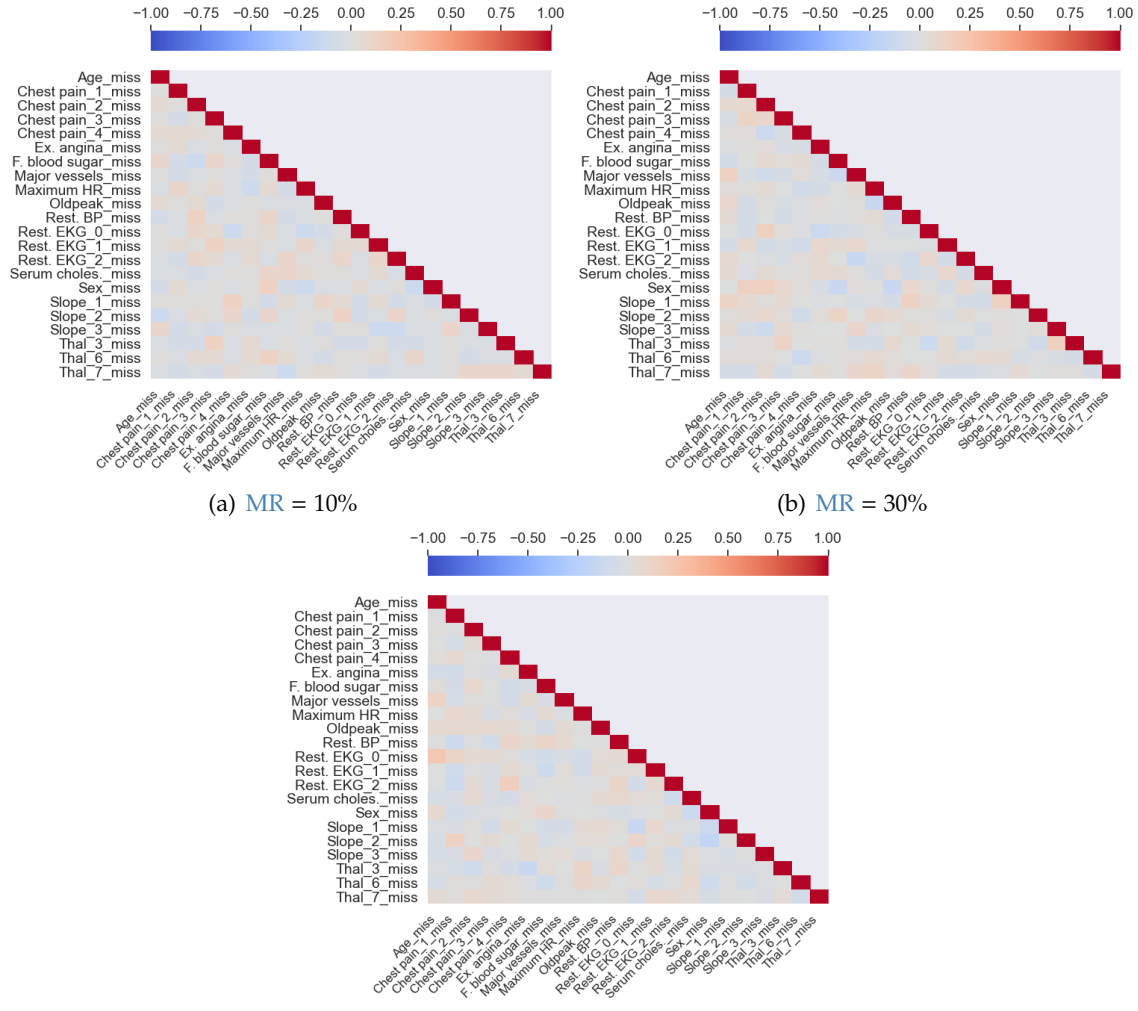
Table 6.5: Average of the standard deviations of the nullity correlation matrices obtained for different MRs under the same missingness mechanism. For legibility purposes, the values are multiplied by 10^2.

	(**Mean ± Standard Deviation**) $\times 10^2$		
Dataset	**MCAR**	**MAR**	**MNAR**
Wine	5.34 ± 2.77	7.91 ± 3.97	13.20 ± 7.55
SPECT Heart	4.96 ± 2.69	12.47 ± 5.69	13.76 ± 5.78
Statlog (Heart)	4.30 ± 2.30	12.11 ± 6.49	9.42 ± 4.47

In every dataset, the average of the standard deviations has similar values for all mechanisms, with MCAR yielding the lowest averages. Furthermore, the disparity between matrices under the same mechanism should be insignificant, considering that correlation varies from -1 to 1 and that all values on Table 6.5 are under 0.15.

A visual inspection of nullity correlation matrices from every missingness mechanism will be performed below. The upper-triangle of the matrices was omitted for legibility purposes, as the information provided is redundant.

The nullity correlation matrices for the Statlog (Heart) Data Set under the MCAR mechanism with multiple MRs are shown in Figure 6.9. Apart from the main diagonal of the matrices, there is no correlation with an absolute value greater than 0.25. Hence, the correlation between missingness patterns is weak or even non-existent, which is expected given the random nature of the MCAR mechanism. This confirms the inferences drawn from Table 6.4. The independence between data and missingness also explains the absence of a noticeable pattern across MRs. The existing variations between values are rather minor, which leads to the lower standard deviations present in Table 6.5. Figure B.8 shows that the SPECT Heart Data Set yields a similar behaviour under the MCAR mechanism. In this case, there is no correlation value with a modulus higher than 0.21 (apart from the main diagonal of the matrix).



(a) MR = 10%

(b) MR = 30%

(c) MR = 50%

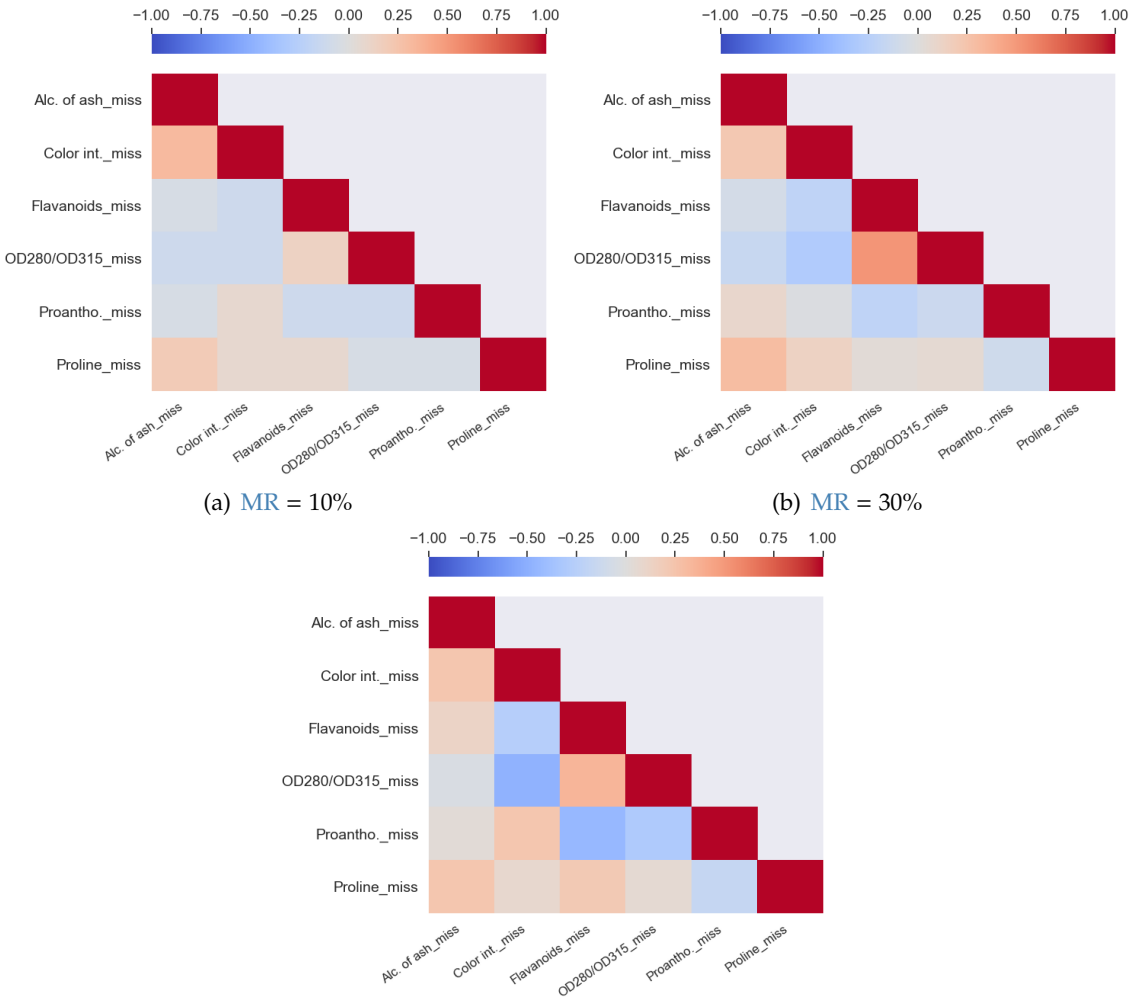
Figure 6.9: Correlation matrices between the missingness patterns of the Statlog (Heart) Data Set obtained for different MRs under the MCAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%.

Figure 6.10 includes the nullity correlation matrices for the Wine Data Set under the MAR mechanism with different MRs. In contrast to the example above, the nullity correlation is moderately strong for higher MRs. A growth in the MR might strengthen existent relationships between missingness and accentuate an underlying pattern across matrices. Note that there are only a few pairs whose nullity correlation has changed significantly in modulus across MRs, which explains why the average standard deviation shown in Table 6.5 is not high. For instance, the negative correlation between the variables "OD280/OD315_miss" and "Color int._miss" has a significant growth (in modulus). Recall that the MAR missingness in these variables is determined by the values of other attributes. This negative nullity correlation may indicate that the values of the determining features of these two missing attributes are negatively correlated. In fact, the correlation matrix of the original Wine Data Set (Figure 6.4(a)) shows that the values of "Hue" and "Malic acid", i.e. the determining features of "OD280/OD315_miss" and "Color int._miss" respectively, have a moderate negative correlation.

Hence, a general inference is drawn: when the missingness patterns of two MAR attributes have a significant positive (negative) correlation, then the values of their determining features have a significant positive (negative) correlation. These moderate nullity correlations are reflected in the higher values yielded by the MAR mechanism in Table 6.4, when compared to MCAR. These conclusions also apply to the nullity correlation matrices of the Statlog (Heart) Data Set under the MAR mechanism, exhibited in Figure B.9. The matrices present an underlying pattern across MRs as well, with the difference that not all nullity correlations that were strong for the lower MR remained so for higher MRs.

Finally, the nullity correlation matrices for the Wine Data Set under the MNAR mechanism with different MRs are shown in Figure 6.11. Similar to the MAR missingness, the MNAR nullity correlations in general increase as the MR rises, reinforcing an underlying pattern. For example, the positive nullity correlation between the variables "OD280/OD315_miss" and "Flavanoids_miss" experiences an overall growth, although it decreases between MR = 30% and MR = 50%. Recall that in every MNAR variable, the missingness is related to the values of the missing variable itself. Hence, the aforementioned positive nullity correlation suggests that the values (observed and unobserved) of "OD280/OD315" and "Flavanoids" are positively correlated. As a matter of fact, the correlation matrix of the original Wine Data Set (Figure 6.4(a)) proves this deduction.

Once again, a general inference is drawn: when the missingness patterns of two MNAR attributes have a significant positive (negative) correlation, then their values, observed and unobserved, have a significant positive (negative) correlation. The occurrence of these correlations explains why the values in Table 6.4 are greater for the MNAR mechanism in comparison to MCAR, where no such significant associations are found. Furthermore, the nullity correlation matrices of the SPECT Heart Data Set under the MNAR mechanism, depicted in Figure B.10, present a similar behaviour to the described above. In this case, there is also an underlying pattern accentuated across MRs, but there is an overall decrease in the nullity correlations as the MR increases.



(a) MR = 10%

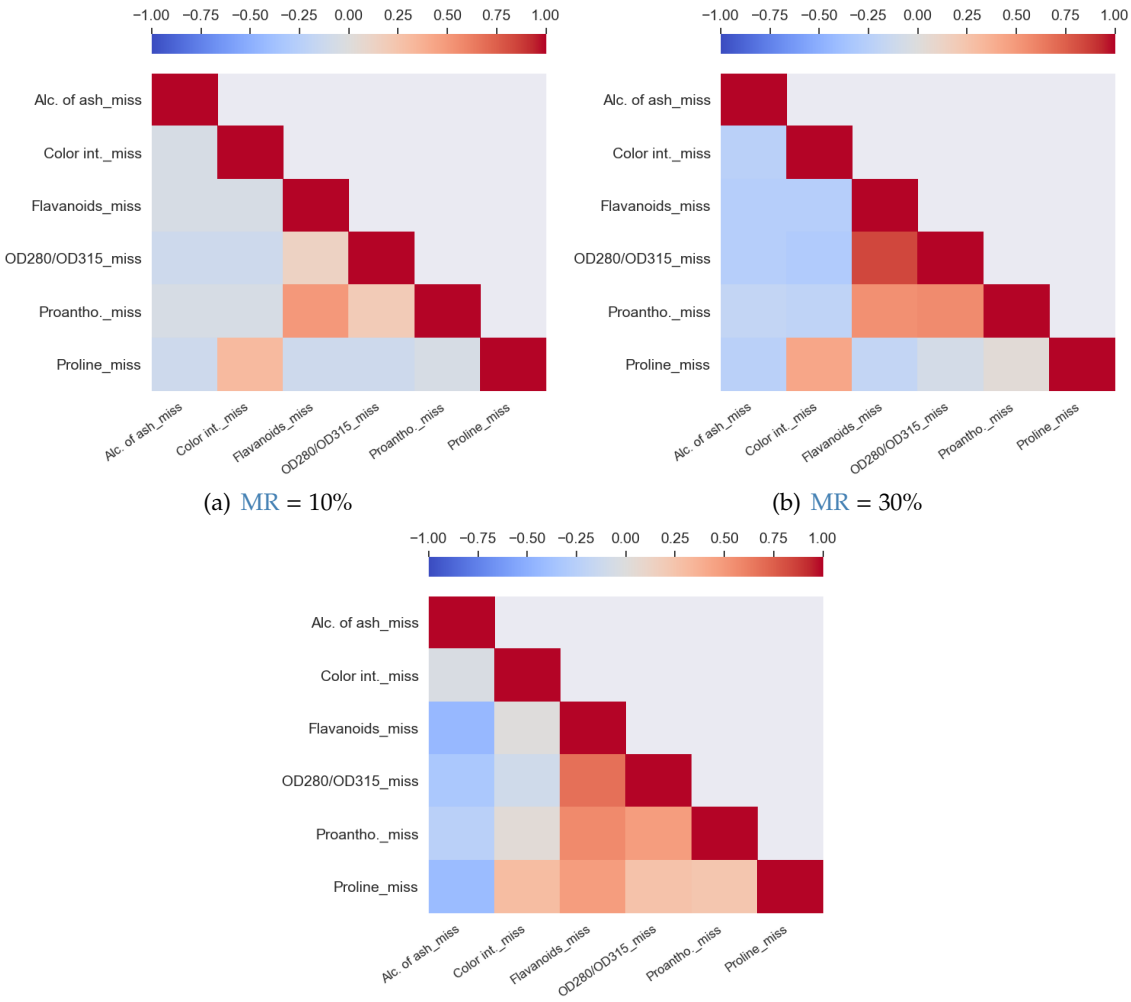
(b) MR = 30%

(c) MR = 50%

Figure 6.10: Correlation matrices between the missingness patterns of the Wine Data Set obtained for different MRs under the MAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%.

In summary, the MCAR mechanism yields inferior nullity correlations than the remaining mechanisms due to the random nature of this missingness, i.e. its independence from measured data. Moreover, there is not a consistent pattern in the nullity correlation matrices across MRs.

On the other hand, MAR and MNAR mechanisms exhibit a pattern that becomes more pronounced as the MR grows. Furthermore, their nullity correlation matrices show the existence of moderate associations, which were expected given the values of Table 6.4. In the MAR case, a significant positive (negative) correlation between the missingness patterns of two attributes implies that the values of their determining features present a significant positive (negative) correlation. As for the MNAR case, a significant positive (negative) correlation between the missingness patterns of two attributes indicates that their values, observed and unobserved, have a significant positive (negative) correlation.



(a) MR = 10% (b) MR = 30%

(c) MR = 50%

Figure 6.11: Correlation matrices between the missingness patterns of the Wine Data Set obtained for different MRs under the MNAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%.

Finally, note that the nullity correlation matrices from different mechanisms are clearly distinct only for higher MRs and even then, the MAR and MNAR matrices are quite similar. Hence, it is impossible to unequivocally identify the missingness mechanisms through nullity correlations alone. Although only specific examples have been discussed, the inferences drawn fairly generalise to any dataset under these three mechanisms.

6.1.5 Final Remarks

This study focused on correlation, a measure of mutual and non-directional association between two variables. There are many other forms of probabilistic dependency that were not considered, such as causality, conditional independence, and multicollinearity.

Although correlation does not ensure causality, Section 6.1.3 showed that it was able to capture the dependency between the missingness of a MAR attribute and the values

of its determinant feature. Furthermore, the associations between variables and MNAR missingness, despite being spurious, may provide useful information. Hence, it is worth investigating if accounting for the correlation between values and missingness pattern has benefits for missing value imputation.

6.2 Evaluation of the Proposed Imputation Methods

In order to evaluate the performance of the proposed imputation methods, a comparative study was conducted, in which existent techniques, both standard and state-of-the-art served as benchmarks. This study encompassed two types of evaluation: quality of imputation evaluation and classification evaluation.

6.2.1 Quality of Imputation Evaluation

Firstly, the quality of the imputation procedure was assessed. Using the synthetic datasets with injected missing values generated from the three UCI Machine Learning Repository databases, the imputed values were compared with the original ones (ground truth), and the precision of the estimation was measured through the MAE.

The results obtained for the proposed methods were compared against seven existing imputation methods: Mean / Mode, KNN, NMVI [86], CMIM [6], CoHiKNN [7], Regression, and MICE [74]. The novel techniques CWKNNI and KNNSCI were both subdivided into four variants, according to the binary values of the parameters **initial_fill** and **update**: the suffix "_FF" corresponds to **initial_fill** = False and **update** = False, "_FT" corresponds to **initial_fill** = False and **update** = True, and so on. This allowed for an analysis of whether there is an optimal combination of these two parameters.

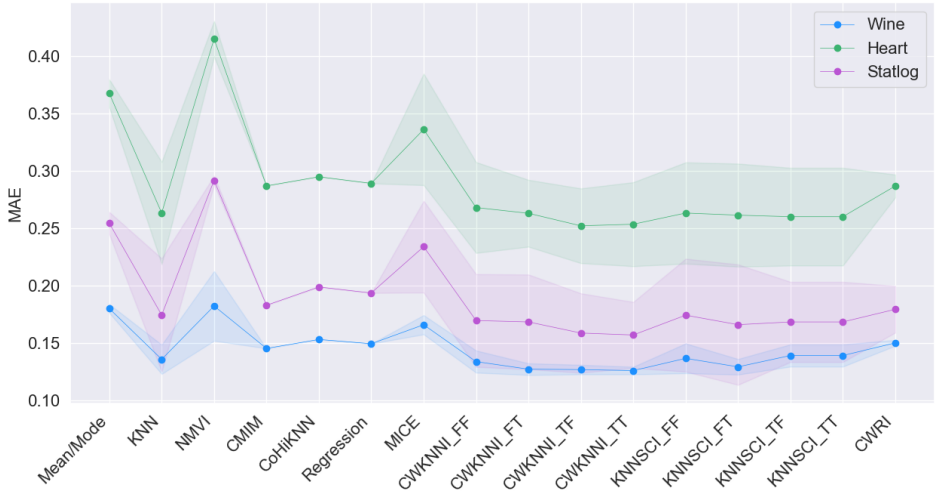
Due to the high quantity of data and the intention to carry out a comprehensive comparative study, it was necessary to aggregate the results. Particularly, the average value of the three distinct MAEs obtained for the three chosen MRs was calculated for every missingness mechanism. Although this does not allow the influence of the MR on the imputation procedure to be assessed, it was deemed more relevant to evaluate the impact of the mechanism.

Table 6.6 contains the average MAEs from all cross validation folds of each imputation method, calculated for every synthetic dataset under the three missingness mechanisms. For each method, only the optimal tested values for their parameters (e.g. number of neighbours) were considered. The highlighted values are the lower MAEs in each column. Techniques that are unable to perform imputation for higher MRs are marked because the calculated MAEs do not fully represent their efficiency. Since this error frequently grows with the MR, because the amount of useful information decreases, then the quality of the imputation performed by those techniques is likely worse than what the displayed MAEs indicate. Figure 6.12 and Figure B.11 provide different visual representations of Table 6.6.

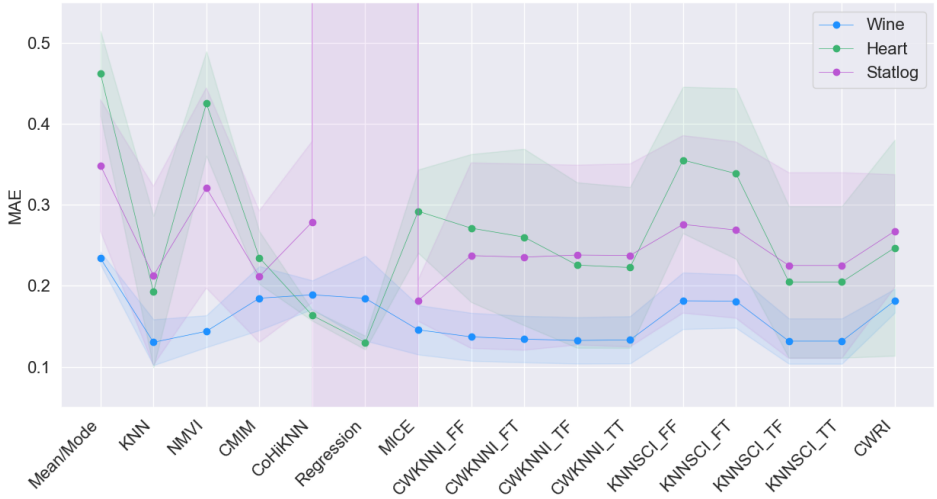
Table 6.6: Average MAE calculated under the three missingness mechanisms. The highlighted values are the lower MAEs in each column.

Method	MAE ± Standard Deviation								
	MCAR			MAR			MNAR		
	Wine	Heart	Statlog	Wine	Heart	Statlog	Wine	Heart	Statlog
Mean / Mode	0.18 ± 0.00	0.37 ± 0.01	0.25 ± 0.01	0.23 ± 0.01	0.46 ± 0.05	0.35 ± 0.08	0.35 ± 0.00	0.62 ± 0.11	0.41 ± 0.13
KNN	0.14 ± 0.01	0.26 ± 0.04	0.17 ± 0.05	**0.13 ± 0.03**	**0.19 ± 0.09**	0.21 ± 0.11	0.24 ± 0.05	0.38 ± 0.17	0.30 ± 0.11
NMVI	0.18 ± 0.03	0.42 ± 0.02	0.29 ± 0.00	0.14 ± 0.02	0.43 ± 0.06	0.32 ± 0.12	**0.16 ± 0.04**	0.49 ± 0.08	0.33 ± 0.05
CMIM	0.15[a]	0.29[a]	0.18[a]	0.19 ± 0.04	0.23 ± 0.03[b]	0.21 ± 0.08	0.27 ± 0.06	0.41 ± 0.18	**0.18 ± 0.02**[b]
CoHiKNN	0.15[a]	0.29[a]	0.20[a]	0.19 ± 0.02	**0.16 ± 0.01**[b]	0.28 ± 0.10	0.29 ± 0.02	0.41 ± 0.27	0.30 ± 0.04[b]
Regression	0.15[a]	0.29[a]	0.19[a]	0.19 ± 0.05	**0.13 ± 0.01**[b]	19.02 ± 20.38	0.27 ± 0.07	**0.24 ± 0.04**[b]	**0.20 ± 0.01**[b]
MICE	0.17 ± 0.01	0.34 ± 0.05	0.23 ± 0.04	0.15 ± 0.03	0.29 ± 0.05	**0.18 ± 0.03**	0.24 ± 0.05	**0.35 ± 0.08**	**0.27 ± 0.03**
CWKNNI_FF	**0.13 ± 0.01**	0.27 ± 0.04	0.17 ± 0.04	0.14 ± 0.03	0.27 ± 0.09	0.24 ± 0.11	0.24 ± 0.05	0.40 ± 0.19	0.31 ± 0.12
CWKNNI_FT	**0.13 ± 0.01**	0.26 ± 0.03	0.17 ± 0.04	**0.13 ± 0.03**	0.26 ± 0.11	0.24 ± 0.12	0.24 ± 0.05	0.39 ± 0.20	0.31 ± 0.12
CWKNNI_TF	**0.13 ± 0.00**	**0.25 ± 0.03**	0.16 ± 0.03	**0.13 ± 0.03**	0.23 ± 0.10	0.24 ± 0.11	0.24 ± 0.05	0.40 ± 0.22	0.31 ± 0.12
CWKNNI_TT	**0.13 ± 0.00**	**0.25 ± 0.04**	0.16 ± 0.03	**0.13 ± 0.03**	0.22 ± 0.10	0.24 ± 0.11	0.24 ± 0.05	0.39 ± 0.23	0.31 ± 0.12
KNNSCI_FF	0.14 ± 0.01	0.26 ± 0.04	0.17 ± 0.05	0.18 ± 0.04	0.36 ± 0.09	0.28 ± 0.11	0.29 ± 0.03	0.40 ± 0.22	0.35 ± 0.10
KNNSCI_FT	**0.13 ± 0.01**	0.26 ± 0.04	0.17 ± 0.05	0.18 ± 0.03	0.34 ± 0.11	0.27 ± 0.11	0.29 ± 0.03	0.39 ± 0.22	0.35 ± 0.10
KNNSCI_TF	0.14 ± 0.01	0.26 ± 0.04	**0.17 ± 0.03**	**0.13 ± 0.03**	0.20 ± 0.09	0.23 ± 0.11	0.24 ± 0.05	0.39 ± 0.16	0.30 ± 0.11
KNNSCI_TT	0.14 ± 0.01	0.26 ± 0.04	0.17 ± 0.03	**0.13 ± 0.03**	0.20 ± 0.09	0.23 ± 0.11	0.24 ± 0.05	0.39 ± 0.16	0.30 ± 0.11
CWRI	0.15 ± 0.00	0.29 ± 0.01	0.18 ± 0.02	0.18 ± 0.02	0.25 ± 0.13	0.27 ± 0.07	0.31 ± 0.01	0.45 ± 0.22	0.36 ± 0.07

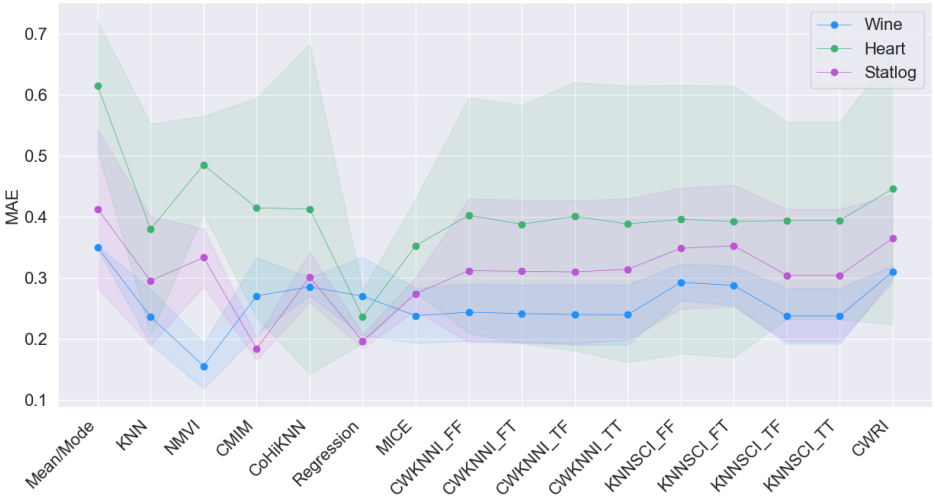
[a] No average was computed and only the MAE for a MR of 10% is shown because the technique was unable to perform imputation for MR = 30% and MR = 50%.
[b] The average was only computed with the MAEs for MRs of 10% and 30% because the technique was unable to perform imputation for MR = 50%.



(a) MCAR



(b) MAR



(c) MNAR

Figure 6.12: Average MAE for all synthetically generated datasets, under each missingness mechanism: (a) MCAR, (b) MAR, (c) MNAR.

Looking at the MCAR mechanism first, the quality of the imputation performed by the proposed methods is similar to the remaining techniques. Overall, CWKNNI yields slightly better results, with the variants CWKNNI_TF and CWKNNI_TT always having the lowest MAEs out of all methods.

Furthermore, Figure 6.12(a) demonstrates that the ranking concerning the quality of the MCAR imputation procedure remains nearly unchanged independently of the dataset, i.e. techniques that are superior (inferior) in the Wine Data Set, are also superior (inferior) in the SPECT Heart Data Set and the Statlog (Heart) Data Set. Hence, for MCAR data of both numeric and binary types, it is inferred that CWKNNI will produce estimates that are closer to the real values if they had been observed.

Figure 6.12(a) also shows that the MAEs are always lower for the Wine Data Set, followed by the Statlog (Heart) Data Set, and then the SPECT Heart Data Set, i.e. the imputation quality is better in the dataset which exclusively included numeric attributes, followed by the mixed-type dataset, and then the dataset with only binary variables. This suggests that the imputation performed by the selected methods is more precise in MCAR numeric data than in MCAR binary values. However, it is important to note that the MAE for binary variables represents the proportion of falsely imputed categories, and therefore has a tendency to be higher than for numeric variables, where disparities between the imputed and original values are less heavily penalised.

As for the MAR mechanism, the proposed methods also exhibit an imputation quality close to their competitors. All four variants of CWKNNI, along with KNNSCI_TF and KNNSCI_TT, are amongst the methods with the highest overall precision.

The ranking concerning the quality of the MAR imputation procedure, as depicted by the three lines plotted in Figure 6.12(b), is not as identical in the three datasets as it was in the case of MCAR. In fact, the only pattern equal in all three lines is the increase of the MAE in the variants KNNSCI_FF and KNNSCI_FT. Furthermore, this graph facilitates the observation that CWKNNI, KNNSCI_TF and KNNSCI_TT provide fairly accurate estimates for the MAR values.

Once again, the MAEs are generally lower for the Wine Data Set. The lines that represent the Statlog (Heart) Data Set and the SPECT Heart Data Set often intersect, and therefore there is not a dataset with a clearly poorer imputation. For the reasons stated above, it is not fair to consider that the selected methods are more accurate in the imputation of numeric values than of binary ones solely based on the MAE.

Recall that the injection of MAR values was based on the pairing of highly correlated features, where one of the features determines the missing elements in the other. Hence it was expected that the proposed methods would yield the best results out of all techniques, since they account for the correlation between values in the imputation process. Although their imputation quality was among the best, this superiority was not observed. The impact of missing values on the correlation between values, which grows with the MR as proved in Section 6.1.2, may have harmed the precision of the imputation procedure.

Finally, the proposed methods show a satisfactory imputation quality in the MNAR

mechanism. Contrary to the mechanisms above, there is not a case where these techniques are superior to the remaining, but the results are nevertheless comparable. Similar to the MAR mechanism, the proposed methods that normally produce produce lower MAEs are CWKNNI (all variants), KNNSCI_TF and KNNSCI_TT.

Even though the methods CMIM, CoHiKNN, and Regression show better MNAR imputation results in the Statlog (Heart) Data Set, their average MAE does not include the error corresponding to MR = 50%, as these methods were unable to perform imputation. Since the MAE frequently increases with the MR, it can be inferred that these techniques would have a poorer performance than the proposed methods had they been able to perform imputation, similar to what is observed in the Wine Data Set. The same deduction can be made for Regression imputation in the SPECT Heart Data Set.

The ranking concerning the quality of the MNAR imputation procedure presents a considerable similarity across the three datasets, as exhibited in Figure 6.12(c). Particularly, the two lines that represent the Wine Data Set and the Statlog (Heart) Data Set, have a nearly equal behaviour regarding the proposed methods. This figure demonstrates that CWKNNI, KNNSCI_TF and KNNSCI_TT yield reasonably accurate MNAR estimates, considering that these techniques are capable of performing imputation on all tested MRs.

Moreover, Figure 6.12(c) shows that the MAEs regarding MNAR imputation are overall lower for the Wine Data Set, followed by the Statlog (Heart) Data Set, and then the SPECT Heart Data Set. The only three techniques where this order does not hold are CMIM, CoHiKNN, and Regression, i.e. the three methods whose average MAE may not fully represent their precision. Once again, no just inference can be drawn regarding the imputation quality on MNAR numeric data versus MNAR binary values.

Additionally, Figure B.11 reveals that, for any dataset, the MNAR mechanism generally presents the largest MAEs in all techniques. This was expected given that most existing techniques, both standard and state-of-the-art, are MAR-based approaches, and thus provide better results under the MCAR and MAR mechanisms. In fact, no significant difference was observed between the MAEs of these two mechanisms, although overall MCAR errors are slightly lower in the Wine Data Set and Statlog (Heart) Data Set. Furthermore, the three lines plotted in each graph of Figure B.11 exhibit a comparable behaviour, although it is not as noticeable as in Figure 6.12(a), for example. This indicates that the ranking concerning the quality of the imputation procedure differs moderately depending on the mechanism, with the most prominent variations occurring in the SPECT Heart Data Set (Figure B.11(b)).

To summarise, the proposed methods yield consistently good results in all three missingness mechanisms. In fact, in comparison to its competitors, CWKNNI is the most precise technique in the MCAR mechanism. The four variants of CWKNNI demonstrate a similar imputation quality. As for KNNSCI, the variants in which **initial_fill** = True show a higher quality in the MAR and MNAR mechanisms than the other two variants. KNNSCI_TF and KNNSCI_TT both provide fairly accurate estimates under all three missingness mechanisms. Furthermore, the imputation performed by the proposed

methods is overall less precise on MNAR data, which is in compliance with the state-of-the-art techniques.

In terms of computational cost, both KNNSCI and CWRI reveal a lower cost than the majority of the implemented state-of-the-art techniques. However, CWKNNI is one of the most expensive methods, which precluded its application on the Cardiothoracic Surgery Dataset, i.e. the most complex working dataset. If the missing value imputation procedure is not a real-time task, a high computational cost should not be a factor that determines the rejection of a certain method, provided that the imputation performed contributes to the improvement of data quality.

Finally, note that this comparative study evaluated the imputation quality of the proposed methods under diverse missingness conditions and on distinct variable types, thus ensuring a comprehensive assessment that is often lacking in the literature.

6.2.2 Classification Evaluation

Since the original values of the missing elements in real-world datasets are not known, it is not possible to assess the quality of the imputation procedure through the MAE or any other measurement that requires a ground truth. Therefore, a classification evaluation was carried out, in which the impact of the imputation procedure on the performance of different ML models was studied. Three ML algorithms were trained upon the imputed datasets: a RF classifier, a SVM classifier, and a NB classifier. Additionally, these classifiers were also trained upon datasets to which listwise deletion was applied.

Similar to Section 6.2.1, it was necessary to aggregate the results of the synthetic datasets due to the high quantity of data. The increased complexity of real-world datasets is reflected in the presence of unknown and possibly multiple missingness mechanisms simultaneously. Hence, it was decided to not evaluate the impact of the missingness mechanism on the AUROCs because such analysis could only be performed on the synthetic datasets, whose imputation quality was already studied quite extensively in the section above. The results regarding the different MRs of the three mechanisms were averaged to reduce the volume of results and thus facilitate their interpretation.

Table 6.7 refers to the RF classifier and contains the average AUROC for every imputation method. Only the optimal values for the parameters of each method (e.g. number of neighbours) were considered. The highlighted values are the highest AUROC scores in each column. The Cardiothoracic Surgery Dataset presents two scores because the information collected in 2019 was assembled in an additional test subset. Techniques that are unable to perform for higher MRs are marked, as the calculated AUROC may not fully represent them. There were also methods that, due to high computational costs, could not conduct imputation on the Cardiothoracic Surgery Dataset, a considerably large and complex dataset. Figure B.12(a) provides a visual representation of Table 6.7.

The majority of the selected methods yields identical results, as reflected in Figure B.12(a), where the plotted lines are nearly constant and the existing variations come

Table 6.7: Average AUROC(%) using a RF classifier. The highlighted values are the higher scores in each column.

Method	(AUROC ± Standard Deviation) %				
	Wine	Heart	Statlog	Osteoporosis	Cardiothoracic Surgery
Complete Dataset	100.00±0.00	81.04±10.47	91.47±3.90	N/A	N/A
Listwise Deletion	79.70±24.26[a]	70.04±11.94[a]	**91.38±1.49**[b]	79.96±5.59	**77.55±14.35**[d] 56.13±7.31
Mean / Mode	98.71±0.94	83.77±8.26	89.18±2.22	83.43±3.31	70.18±4.66 61.24±0.97
KNN	98.45±1.07	81.12±7.12	89.05±2.62	83.45±3.20	71.08±4.94 61.17±1.81
NMVI	98.27±2.05	83.86±6.03	88.06±2.59	78.04±2.59	65.61±5.04 57.07±0.94
CMIM	**99.24±0.39**[a]	**86.50±5.50**[a]	89.23±1.68[b]	83.36±3.36	[c]
CoHiKNN	98.85±0.51[a]	**88.93±7.27**[a]	89.30±1.44[b]	83.57±3.12	[c]
Regression	98.90±0.50[a]	81.91±5.03[a]	88.11±3.22[b]	83.21±3.20	70.36±4.81 60.16±0.97
MICE	**98.99±0.69**	83.02±4.83	89.12±1.96	83.37±2.92	70.63±4.51 59.60±1.41
CWKNNI_FF	98.56±0.91	82.05±6.45	89.33±1.71	83.54±2.90	[c]
CWKNNI_FT	98.55±1.05	82.12±5.45	89.09±2.16	83.60±3.04	[c]
CWKNNI_TF	98.51±0.91	82.25±5.19	89.26±1.80	83.55±3.02	[c]
CWKNNI_TT	98.43±1.05	82.04±5.52	89.29±1.50	83.64±2.85	[c]
KNNSCI_FF	98.47±1.15	83.44±5.98	88.88±3.00	83.63±3.07	**71.32±4.78** 61.42±1.53
KNNSCI_FT	98.57±0.98	**84.63±5.58**	89.09±2.07	**83.73±3.13**	71.10±4.80 **61.86±1.00**
KNNSCI_TF	98.45±1.04	81.81±5.89	89.04±2.28	83.11±3.02	70.84±5.31 60.75±2.06
KNNSCI_TT	98.45±1.04	81.81±5.89	89.04±2.28	83.11±3.02	70.84±5.31 60.75±2.06
CWRI	98.72±0.56	83.12±6.26	**89.50±2.26**	83.07±3.66	71.10±1.92 60.23±1.39

[a] The technique was unable to perform under the MCAR mechanism with MR = 30% and MR = 50%.
[b] The technique was unable to perform under the MCAR mechanism with MR = 30% and MR = 50%, and under the MNAR mechanism with MR = 50%.
[c] The technique was unable to perform imputation due to high computational costs.
[d] Only 7.3% of the samples from the Cardiothoracic Surgery Dataset were used in this complete-case analysis, which is not at all representative.

from methods that were unable to perform for certain MRs. The proposed methods are in compliance with their competitors, sometimes even exhibiting higher AUROCs. Table 6.7 shows that, if listwise deletion is not considered, KNNSCI_FT outperforms the other techniques on 3 of the 5 datasets, and CWRI produces the best AUROC in the Statlog (Heart) Data Set. These methods were not particularly superior in terms of imputation quality, which may suggest that a more precise imputation does not imply a better classification.

In order to further investigate this hypothesis, a study on whether an optimal imputation quality, i.e. a null MAE, leads to a better classification model was conducted. To this end, the results from the original (and complete) UCI Machine Learning Repository datasets were compared with those of the imputation methods. In the SPECT Heart Data Set, the vast majority of the classifiers trained upon imputed datasets outperforms the model trained upon the original dataset, thus demonstrating that there is not a clear relationship between imputation quality and the performance of a ML model. This finding raises the question of whether a more suitable imputation method for a DSS is one that yields more precise estimates or one by which a better classification performance is achieved. At first glance, a clinical prediction model should have an optimal performance, but it may not be acceptable to fully disregard the quality of the imputation performed by the chosen method. For instance, consider an imputation technique that produces biased parameter estimates, i.e. distorts the original statistical distribution of the data. In some cases, this permits a greater generalization ability of the ML model that was trained upon the imputed data. However, the knowledge that should have been learned from the data may have been corrupted by the imputation procedure, ultimately leading to a facilitated learning task and misleadingly better classification results.

As for the SVM classifier, Table 6.8 includes the average AUROC of all stratified folds for every imputation method. Once again, only the optimal values for the parameters of each method were considered. The Cardiothoracic Surgery Dataset presents two AUROCs because it has an additional test subset. The highlighted values are the highest AUROCs in each column, and the techniques that were unable to perform under certain conditions were marked. A visual representation of Table 6.8 is depicted in Figure B.12(b).

Similar to the previous case, practically all imputation techniques show nearly equal AUROCs, with the most prominent difference being from the score produced by Listwise Deletion in comparison to the others. As already discussed, this technique should be used cautiously, since simply discarding all incomplete instances may produce a dataset that is not representative of the original problem, particularly for higher MRs.

Table 6.8 reveals that the proposed methods achieved results comparable to the others. Moreover, if the techniques that could not perform under all circumstances are set aside, CWRI yields the highest AUROC in 2 of the 5 datasets, and KNNSCI_FT outperforms its competitors on the Osteoporosis Dataset. These were the methods that also exhibited a better performance for the RF classifier. Once more, given that neither CWRI nor KNNSCI_FT showed a clear greater imputation quality (rather the opposite even), this reinforces that a less precise imputation does not lead to worse classification results. In

Table 6.8: Average AUROC(%) using a SVM classifier. The highlighted values are the higher scores in each column.

Method	(AUROC ± Standard Deviation) %				
	Wine	Heart	Statlog	Osteoporosis	Cardiothoracic Surgery
Complete Dataset	100.00±0.00	83.39±9.43	91.17±3.17	N/A	N/A
Listwise Deletion	91.97±15.99[a]	58.39±24.53[a]	**89.55±2.97**[b]	82.58±4.33	46.37±18.97 49.00±8.66
Mean / Mode	99.27±0.60	75.71±24.30	88.37±1.95	82.78±2.40	68.58±2.89 55.95±2.55
KNN	99.27±0.62	79.44±13.16	88.27±2.32	82.75±2.35	68.88±2.70 55.97±2.44
NMVI	98.78±1.72	**84.72±4.36**	88.31±2.09	80.39±3.19	68.54±2.92 55.61±2.56
CMIM	**99.54±0.34**[a]	**85.42±3.88**[a]	88.81±1.35[b]	82.57±2.52	[c]
CoHiKNN	**99.50±0.36**[a]	**85.47±5.84**[a]	**89.26±1.09**[b]	82.74±2.32	[c]
Regression	**99.41±0.49**[a]	83.85±3.68[a]	87.39±3.90[b]	82.29±2.25	63.72±4.18 **56.67±1.64**
MICE	99.35±0.54	80.77±14.89	88.49±2.00	82.62±2.59	**68.96±2.54** 55.56±2.89
CWKNNI_FF	99.35±0.50	79.80±14.68	88.36±2.32	82.78±2.41	[c]
CWKNNI_FT	99.28±0.56	80.91±9.92	88.24±2.01	82.78±2.41	[c]
CWKNNI_TF	99.25±0.58	80.65±10.36	88.52±1.37	82.91±2.70	[c]
CWKNNI_TT	99.21±0.66	80.72±9.57	88.54±1.46	82.73±2.56	[c]
KNNSCI_FF	99.23±0.67	80.88±13.99	88.35±2.45	82.83±2.49	68.90±2.69 55.98±2.63
KNNSCI_FT	99.19±0.70	83.35±6.00	88.27±1.78	**83.02±2.58**	68.91±2.68 56.14±2.63
KNNSCI_TF	99.23±0.59	77.14±17.41	88.24±2.02	82.71±2.32	68.87±2.72 56.03±2.56
KNNSCI_TT	99.23±0.59	77.14±17.41	88.24±2.02	82.71±2.32	68.87±2.72 56.03±2.56
CWRI	**99.40±0.44**	84.32±5.76	**88.91±1.84**	82.78±2.41	68.58±1.85 55.93±3.24

[a] The technique was unable to perform under the MCAR mechanism with MR = 30% and MR = 50%.
[b] The technique was unable to perform under the MCAR mechanism with MR = 30% and MR = 50%, and under the MNAR mechanism with MR = 50%.
[c] The technique was unable to perform imputation due to high computational costs.

fact, the SPECT Heart Data Set has cases where the AUROC of classifiers trained upon imputed datasets is greater than the one from the model trained upon the original dataset.

Finally, Table 6.9 concerns the NB classifier and displays the average AUROC of all folds for every imputation method. The results were selected and presented in a similar manner to the last two tables. Figure B.12(c) gives a visual representation of Table 6.9.

As in previous classifiers, most imputation techniques produce almost identical results in all working datasets, which is reflected in the nearly unchanging lines plotted in the graph of Figure B.12(c). The proposed methods are in compliance with the remaining ones, and can even be considered superior in some datasets. Particularly, out of the techniques that were able to perform under all circumstances, the variants CWKNNI_TF and CWKNNI_FT respectively exhibit the best AUROC in the SPECT Heart Data Set and in the Osteoporosis Dataset. These methods were amongst the ones with the highest overall imputation precision. However, this does not signify that a superior imputation quality leads to better classification. In fact, beyond the examples already discussed above, Table 6.9 shows that the SPECT Heart Data Set and the Statlog (Heart) Data Set present a few cases in which the AUROC of the NB classifier trained upon an imputed dataset is higher than the one from the model trained upon the complete dataset.

In conclusion, the proposed correlation-based methods yield similar results to state-of-the-art techniques in all five working datasets, regardless of the classifier. There were even some cases where the proposed methods were superior to their competitors. However, the AUROCs were not substantially different from the scores of simpler techniques such as mean / mode imputation, which revealed an overall worse imputation quality. In fact, it was observed that a more precise imputation does not necessarily generate better classification results. Note that this classification evaluation took into account not only the AUROCs, but also whether the methods could perform under all the circumstances tested. Lastly, no imputation technique was clearly favoured by a certain classifier, as the results were fairly consistent in all three ML models. It was observed that there are more suitable classifiers for certain problems than others, which is natural given that distinct ML algorithms have different approaches regarding the processing and evaluation of data.

For each working dataset, Table 6.10 summarises the highest average AUROC of every classifier, detailing the corresponding imputation method and additional performance metrics. The imputation techniques that were unable to perform for higher MRs were not considered. No threshold optimisation was performed, and therefore the default value of 0.5 was considered for interpreting probabilities to class labels. A threshold optimisation would ensure that the best classification metrics were achieved for each model. However, it is not expected for this procedure to significantly change the results presented given that, during the training of the classifiers, the weights associated to each class were required to be inversely proportional to its frequency, thus addressing imbalanced classes. Table B.1 specifies the values of the hyperparameters and parameters of the classifiers and imputation methods, respectively, which were shown in Table 6.10.

Table 6.9: Average AUROC using a NB classifier. The highlighted values are the higher scores in each column.

Method	(AUROC ± Standard Deviation) %				
	Wine	Heart	Statlog	Osteoporosis	Cardiothoracic Surgery
Complete Dataset	99.82±0.25	74.86±4.86	86.50±4.33	N/A	N/A
Listwise Deletion	82.30±20.3[a]	68.97±9.65[a]	82.12±6.69[b]	77.28±5.35	50.85±14.71 49.58±3.16
Mean / Mode	97.79±1.90	74.63±5.95	85.50±1.55	77.54±2.82	66.58±4.57 58.00±0.88
KNN	98.21±1.13	75.92±4.04	85.93±1.99	77.80±2.93	67.05±4.27 57.86±0.78
NMVI	97.57±2.21	74.32±3.55	86.07±1.81	77.54±2.62	66.90±4.36 58.00±0.85
CMIM	98.07±1.48[a]	**77.97±5.04**[a]	85.32±1.42[b]	77.47±2.60	[c]
CoHiKNN	**98.58±0.81**[a]	**78.01±5.85**[a]	85.84±0.61[b]	77.58±2.93	[c]
Regression	98.00±1.50[a]	74.45±3.16[a]	85.69±2.37[b]	76.51±2.56	50.80±7.18 49.56±3.97
MICE	**98.45±1.04**	75.44±2.72	**87.38±1.66**	77.59±2.69	**67.45±4.07** **58.04±1.15**
CWKNNI_FF	98.28±1.08	75.55±6.13	86.42±1.40	77.78±2.93	[c]
CWKNNI_FT	98.19±1.10	76.35±3.92	86.22±1.91	**77.90±2.89**	[c]
CWKNNI_TF	98.16±1.10	**77.36±5.70**	86.63±0.91	77.77±2.94	[c]
CWKNNI_TT	98.12±1.13	76.46±5.07	86.42±1.23	77.83±2.89	[c]
KNNSCI_FF	98.28±1.10	76.60±3.74	85.99±2.33	77.81±2.93	67.00±4.25 57.94±0.78
KNNSCI_FT	98.30±1.15	75.44±4.15	86.15±1.93	77.89±2.87	67.06±4.26 57.94±0.79
KNNSCI_TF	98.28±0.95	76.41±4.29	86.19±1.46	77.79±3.05	67.04±4.28 57.86±0.78
KNNSCI_TT	98.28±0.95	76.41±4.29	86.19±1.46	77.79±3.05	67.04±4.28 57.86±0.78
CWRI	98.12±1.40	75.11±5.13	86.28±1.57	77.55±2.80	65.07±3.50 57.56±0.68

[a] The technique was unable to perform under the MCAR mechanism with MR = 30% and MR = 50%.
[b] The technique was unable to perform under the MCAR mechanism with MR = 30% and MR = 50%, and under the MNAR mechanism with MR = 50%.
[c] The technique was unable to perform imputation due to high computational costs.

Table 6.10: Highest average AUROC(%) of every classifier, corresponding imputation model and additional performance metrics. (Mean ± Standard Deviation).

	Classifier	Imputation Method	AUROC (%)	Accuracy (%)	Sensitivity (%)	Precision (%)	F1-Score (%)	Specificity (%)
Wine	RF	MICE	98.99 ± 0.69	92.47 ± 2.58	93.06 ± 2.48	93.04 ± 2.24	92.52 ± 2.52	96.29 ± 1.30
	SVM	CWRI	99.40 ± 0.44	93.00 ± 2.34	93.58 ± 2.18	93.56 ± 2.08	93.10 ± 2.67	96.51 ± 1.20
	NB	MICE	98.45 ± 1.04	90.51 ± 3.87	91.03 ± 3.67	91.77 ± 3.29	90.69 ± 3.83	95.15 ± 1.99
Heart	RF	KNNSCI_FT	84.63 ± 5.58	84.28 ± 3.06	71.85 ± 6.93	69.30 ± 5.38	66.38 ± 5.99	57.04 ± 11.81
	SVM	NMVI	84.72 ± 4.36	75.56 ± 5.06	73.87 ± 2.51	62.03 ± 3.67	60.57 ± 4.88	71.85 ± 5.24
	NB	CWKNNI_TF	77.36 ± 5.70	67.12 ± 5.87	69.28 ± 5.54	57.99 ± 4.32	53.54 ± 6.01	71.85 ± 8.18
Statlog	RF	CWRI	89.50 ± 2.26	81.40 ± 3.26	81.21 ± 3.22	81.54 ± 3.27	81.12 ± 3.30	82.93 ± 4.29
	SVM	CWRI	88.91 ± 1.84	81.84 ± 2.37	81.64 ± 2.25	81.89 ± 2.48	81.59 ± 2.36	83.44 ± 3.74
	NB	MICE	87.38 ± 1.66	80.70 ± 2.18	80.47 ± 2.24	80.76 ± 2.25	80.43 ± 2.23	82.52 ± 1.93
Osteoporosis	RF	KNNSCI_FT	83.73 ± 3.13	75.65 ± 2.49	75.34 ± 2.27	74.65 ± 2.46	74.80 ± 2.43	73.93 ± 2.77
	SVM	KNNSCI_FT	83.02 ± 2.58	75.41 ± 2.39	75.31 ± 2.90	74.43 ± 2.57	74.62 ± 2.62	74.88 ± 5.56
	NB	CWKNNI_FT	77.90 ± 2.89	72.98 ± 3.62	70.22 ± 4.10	71.64 ± 3.97	70.63 ± 4.09	57.94 ± 6.61
Cardiothoracic Surgery	RF	KNNSCI_FF	71.32 ± 4.78	82.65 ± 2.98	63.16 ± 4.07	56.89 ± 2.51	57.89 ± 3.44	85.95 ± 3.08
		KNNSCI_FT	61.86 ± 1.00	84.17 ± 1.87	56.32 ± 0.86	54.16 ± 0.49	54.58 ± 0.51	88.96 ± 2.31
	SVM	MICE	68.96 ± 2.54	76.08 ± 2.42	63.17 ± 2.91	52.91 ± 0.93	54.21 ± 1.70	78.31 ± 2.52
		Regression	56.67 ± 1.64	77.87 ± 0.96	55.18 ± 1.73	52.29 ± 0.76	51.59 ± 1.01	82.78 ± 1.09
	NB	MICE	67.45 ± 4.07	83.84 ± 0.74	62.97 ± 2.01	57.05 ± 1.02	58.31 ± 1.22	87.38 ± 0.98
		MICE	58.04 ± 1.15	85.22 ± 0.31	54.63 ± 0.50	53.45 ± 0.41	53.82 ± 0.47	90.49 ± 0.30

Chapter 7

Conclusions and Future Work

This chapter concludes this dissertation by reviewing its main findings and contributions to biomedical research. Furthermore, a reflection on this work's limitations is presented, along with perspectives for future approaches to address the challenge of missing data.

7.1 Conclusions

Missing data are ubiquitous in biomedical sciences, posing a recurring predicament to delivering reliable AI-based DSS. There has been a growing interest around strategies that address this inevitable challenge, specifically missing value imputation techniques.

Despite a noticeable research endeavour, state-of-the-art imputation methods still face drawbacks that prevent the forthcoming DSSs' application. Furthermore, one of the most widely-used approaches is a simple complete-case analysis, which handles missing values as a disposable nuisance, occasionally yielding dubious results.

This dissertation proposed novel imputation techniques that not only sought to tackle the issues imposed by missing data, but also overcame some of the limitations presented by existing methods. Moreover, the adopted approach seeked to leverage correlation as a potential tool for missing value imputation. Even though several authors have shown the benefits of this strategy in distinct fields, the exploration of correlation within the scope of missing value imputation is still overlooked in clinical data.

Hence, the first stage of this work contemplated a fairly thorough correlational study, i.e. an analysis of the relationships between variables, covering every missingness mechanism fully established in the literature. Instead of focusing solely on the correlation between the values of two attributes as most studies do, this work also encompassed the correlation between values and missingness pattern, and the correlation between the missingness patterns of two distinct variables. The choice to analyse these types of correlation stemmed

from the fact that each missingness source mechanism is characterised by a different relationship between data and missingness, and therefore it was worth observing how correlation captured these associations.

In regards to **RQ 1.2**, it was concluded that missingness, particularly MNAR, has a significant impact on the correlation between values, which intensified as the MR rose. As for the other types of correlation, the MR overall slightly accentuated existing relationships between values and missingness patterns or between two missingness patterns. When developing correlation-based imputation methods, it should be assessed whether it is worth considering this effect, specially if the working datasets have higher MRs.

Furthermore, this work concluded that it is not possible to distinguish between the three missingness mechanisms through correlation alone, thus answering **RQ 1.1**. Although the MCAR mechanism has distinct characteristics under certain circumstances, the practical issue with MAR and MNAR remains, i.e. there is no way to verify if the missingness is dependent on unobserved data or is solely related to other measured variables. This might raise the question of whether the current taxonomy concerning missingness mechanisms is helpful for missing data imputation in real-world problems.

Nevertheless, the correlational study permitted a more comprehensive view of the missingness mechanisms and the generic relationships they describe. In particular, the dependency between missingness and observed values provided interesting insights, which inspired the development of three innovative methods that incorporate this association within the imputation procedure: CWKNNI, KNNSCI, and CWRI.

In terms of imputation precision, the proposed correlation-based methods yielded similar and, in some cases, slightly better MAEs than competing state-of-the-art techniques. It is not possible to specify an optimal imputation strategy for each missingness mechanism, since none of the evaluated techniques was consistently superior to the others. Even so, the proposed methods showed that accounting for correlation can enhance the imputation quality in some datasets, particularly under MCAR missingness, which answers **RQ 2.1**.

Afterwards, in order to conduct a classification evaluation, three ML models were trained upon the imputed datasets and their performance was assessed through the AUROC metric. Contrary to the expected, a more precise imputation was not always followed by better classification results, thus covering **RQ 2.3**. This shows that, when developing a robust AI-based DSS, the evaluation of the imputation procedure must not rely merely on the differences between estimated and original values, possibly resorting to domain knowledge in a data-centric approach.

Finally, regardless of the classifier, the proposed correlation-based methods were in compliance with their competitors, sometimes even exhibiting higher AUROC scores. In fact, the developed methods often showed the best AUROCs within the selected real-world case studies. Hence, imputation based on correlation can yield better classification results in real-world medical problems, which answers **RQ 2.2**. In summary, this work confirmed the auspicious role of correlation-based imputation towards the improvement of DSSs' robustness to missing values, while addressing limitations of current imputation methods.

7.2 Limitations and Future Work

The previous section discussed the main contributions of this dissertation, which focused on the exploration of the concept of correlation and its application in missing data imputation methods. Even so, this work presents some limitations that might be addressed in future projects within the scope of delivering reliable AI-based DSSs.

One of the main fragilities is the high computational cost yielded by CWKNNI, one of the proposed missing value imputation methods. Due to time constraints, this method could neither be applied nor evaluated on the most complex working dataset. Note that this limitation is shared with several state-of-the-art methods. Even though computational cost should not be a determining factor for the dismissal of an imputation method, it is a concern that must be taken into account in future works.

Additionally, the optimal values for the parameters **percentage** and **n_neighbors** of the imputation methods CWKNNI and KNNSCI were found through a grid search strategy, i.e. all possible combinations of parameter values were evaluated and the best one was chosen. This approach increased the computational cost, and so an alternative that identifies the optimal values in a more automatic manner would be beneficial.

In regards to data types, this dissertation only worked with structured data, particularly numeric and binary variables. However, clinical data is highly multimodal, encompassing both structured and unstructured data produced from multiple sources and that may carry different physiological information. Multimodal ML is able to cope with this heterogeneity, and may even manage to establish dependencies within and between modalities. Since such dependencies can be leveraged within the scope of missing data handling, it is worth investigating the potential of multimodal AI.

Moreover, this work focused solely on correlation, which is a measure of mutual and non-directional association between two attributes. Although leveraging correlation has revealed to be a promising approach, there are many other forms of probabilistic dependency that were not considered. For instance, causality and conditional independence are two examples of directional associations. Besides, multicollinearity is a statistical measurement of the linear relation among two or more variables. Given the complexity inherent to clinical data, it is expectable that attributes exhibit a wide range of probabilistic dependencies, whose contribution to the processing of missing values should be studied.

Furthermore, in a real-world setting, resorting to domain knowledge to infer about the characteristics of the data or its source might lead to more precise imputation estimates. A future project could consist of analysing how the integration of domain knowledge affects the performance of an imputation approach exclusively based on measured data.

Lastly, given that imputation procedures aim to estimate often unknown values, uncertainty can be explored in order to increase professionals' trust in an AI-based DSS. These systems should not only have an optimised performance, but also provide a well-founded decision and point out possible frailties, such as the number of imputations made and the confidence in each missing value prediction.

Bibliography

[1] J. M. Lourenço, *The NOVAthesis LaTeX Template User's Manual*, NOVA University Lisbon, 2021. [Online]. Available: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (cit. on p. ii).

[2] E. P. Ambinder, "Electronic Health Records", *Journal of oncology practice*, vol. 1, no. 2, p. 57, 2005. DOI: 10.1200/jop.2005.1.2.57 (cit. on pp. 1, 5).

[3] A. Bagheri, T. K. J. Groenhof, W. B. Veldhuis, P. A. de Jong, F. W. Asselbergs, and D. L. Oberski, *Multimodal Learning for Cardiovascular Risk Prediction using EHR Data*, 2020. DOI: 10.48550/ARXIV.2008.11979. [Online]. Available: https://arxiv.org/abs/2008.11979 (cit. on p. 1).

[4] M. Kang and J. Tian, "Machine Learning: Data Pre-processing", *Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things*, pp. 111–130, 2018. DOI: 10.1002/9781119515326.ch5 (cit. on pp. 1, 14, 16).

[5] A. Iranfar, A. Arza, and D. Atienza, *ReLearn: A Robust Machine Learning Framework in Presence of Missing Data for Multimodal Stress Detection from Physiological Signals*, 2021. DOI: 10.48550/ARXIV.2104.14278. [Online]. Available: https://arxiv.org/abs/2104.14278 (cit. on pp. 1, 37, 39).

[6] A. M. Sefidian and N. Daneshpour, "Estimating missing data using novel correlation maximization based methods", *Applied Soft Computing*, vol. 91, p. 106 249, 2020. DOI: 10.1016/j.asoc.2020.106249 (cit. on pp. 2, 6, 23, 32, 33, 35, 39, 54, 82).

[7] X. Liu, X. Lai, and L. Zhang, "A Hierarchical Missing Value Imputation Method by Correlation-Based K-Nearest Neighbors", in *Proceedings of SAI Intelligent Systems Conference*, Springer, 2019, pp. 486–496. DOI: 10.1007/978-3-030-29516-5_38 (cit. on pp. 2, 6, 33, 36, 39, 54, 82).

[8] S. C. Denaxas and K. I. Morley, "Big biomedical data and cardiovascular disease research: Opportunities and challenges", *European Heart Journal-Quality of Care and Clinical Outcomes*, vol. 1, no. 1, pp. 9–16, 2015. DOI: 10.1093/ehjqcco/qcv005 (cit. on pp. 5, 6).

[9] C. A. Caligtan and P. C. Dykes, "Electronic health records and personal health records", in *Seminars in oncology nursing*, Elsevier, vol. 27, 2011, pp. 218–228. DOI: 10.1016/j.soncn.2011.04.007 (cit. on p. 5).

[10] Z. Liu, J. Zhang, Y. Hou, X. Zhang, G. Li, and Y. Xiang, *Machine learning for multimodal electronic health records-based research: Challenges and perspectives*, 2021. DOI: 10.48550/ARXIV.2111.04898. [Online]. Available: https://arxiv.org/abs/2111.04898 (cit. on p. 6).

[11] A. M. Fjell, K. B. Walhovd, C. Fennema-Notestine, *et al.*, "CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease", *Journal of Neuroscience*, vol. 30, no. 6, pp. 2088–2101, 2010. DOI: 10.1523/JNEUROSCI.3785-09.2010 (cit. on p. 6).

[12] E. Sylvestre, G. Bouzillé, E. Chazard, C. His-Mahier, C. Riou, and M. Cuggia, "Combining information from a clinical data warehouse and a pharmaceutical database to generate a framework to detect comorbidities in electronic health records", *BMC medical informatics and decision making*, vol. 18, no. 1, pp. 1–8, 2018. DOI: 10.1186/s12911-018-0586-x (cit. on p. 6).

[13] P. S. Pillai and T.-Y. Leong, *Knowledge-driven generative subspaces for modeling multi-view dependencies in medical data*, 2018. DOI: 10.48550/ARXIV.1812.00509. [Online]. Available: https://arxiv.org/abs/1812.00509 (cit. on p. 6).

[14] F. Liu, L. Zhou, C. Shen, and J. Yin, "Multiple Kernel Learning in the Primal for Multimodal Alzheimer's Disease Classification", *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 3, pp. 984–990, 2013. DOI: 10.1109/JBHI.2013.2285378 (cit. on p. 6).

[15] M. G. Rahman and M. Z. Islam, "FIMUS: a framework for imputing missing values using co-appearance, correlation and similarity analysis", *Knowledge-Based Systems*, vol. 56, pp. 311–327, 2014. DOI: 10.1016/j.knosys.2013.12.005 (cit. on p. 6).

[16] X. Chen, Z. Wei, Z. Li, J. Liang, Y. Cai, and B. Zhang, "Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation", *Knowledge-Based Systems*, vol. 132, pp. 249–262, 2017. DOI: 10.1016/j.knosys.2017.06.010 (cit. on p. 6).

[17] P. Schober, C. Boer, and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation", *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018. DOI: 10.1213/ANE.0000000000002864 (cit. on pp. 6, 7, 9, 10, 59).

[18] D. Kornbrot, "Correlation", in *Encyclopedia of Statistics in Behavioral Science - Volume 1*, B. S. Everitt and D. C. Howell, Eds., John Wiley & Sons, Ltd, 2005, pp. 398–400, ISBN: 978-0-470-86080-9 (cit. on pp. 7, 9, 11).

[19] D. C. Howell, *Statistical Methods for Psychology*. Cengage Learning, 2010, ISBN: 978-1111835484 (cit. on pp. 7, 12).

[20] K. L. Wuensch, "Scales of Measurement", in *International Encyclopedia of Statistical Science*, Springer, 2011, pp. 1283–1285, ISBN: 978-3-642-04898-2 (cit. on pp. 7, 8).

[21] R. Somun-Kapetanović, "Variables", in *International Encyclopedia of Statistical Science*, Springer, 2011, pp. 1639–1640, ISBN: 978-3-642-04898-2 (cit. on p. 7).

[22] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research", *Malawi medical journal*, vol. 24, no. 3, pp. 69–71, 2012 (cit. on p. 9).

[23] H. Akoglu, "User's guide to correlation coefficients", *Turkish journal of emergency medicine*, vol. 18, no. 3, pp. 91–93, 2018. DOI: 10.1016/j.tjem.2018.08.001 (cit. on p. 11).

[24] A. Agresti, *Analysis of Ordinal Categorical Data*. John Wiley & Sons, 2010, vol. 656, ISBN: 9780470594001 (cit. on p. 11).

[25] R. N. Forthofer, E. S. Lee, and M. Hernandez, "3 - Descriptive Methods", in *Biostatistics (Second Edition)*, R. N. Forthofer, E. S. Lee, and M. Hernandez, Eds., Second Edition, San Diego: Academic Press, 2007, pp. 21–69, ISBN: 978-0-12-369492-8. DOI: 10.1016/B978-0-12-369492-8.50008-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012369492850008X (cit. on p. 11).

[26] G. A. Fredricks and R. B. Nelsen, "On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables", *Journal of Statistical Planning and Inference*, vol. 137, no. 7, pp. 2143–2150, 2007, ISSN: 0378-3758. DOI: 10.1016/j.jspi.2006.06.045 (cit. on p. 11).

[27] M.-T. Puth, M. Neuhäuser, and G. D. Ruxton, "Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits", *Animal Behaviour*, vol. 102, pp. 77–84, 2015, ISSN: 0003-3472. DOI: 10.1016/j.anbehav.2015.01.010 (cit. on p. 11).

[28] J. D. Brown, "Point-biserial correlation coefficients", *Statistics*, vol. 5, no. 3, pp. 12–6, 2001 (cit. on p. 11).

[29] S. G. Liao, Y. Lin, D. D. Kang, *et al.*, "Missing value imputation in high-dimensional phenomic data: Imputable or not, and how?", *BMC bioinformatics*, vol. 15, no. 1, pp. 1–12, 2014. DOI: 10.1186/s12859-014-0346-6 (cit. on p. 12).

[30] I. Olkin and R. F. Tate, "Multivariate Correlation Models with Mixed Discrete and Continuous Variables", *The Annals of Mathematical Statistics*, pp. 448–465, 1961. DOI: 10.1214/aoms/1177705052 (cit. on p. 12).

[31] W. Bergsma, "A bias-correction for Cramér's V and Tschuprow's T", *Journal of the Korean Statistical Society*, vol. 42, no. 3, pp. 323–328, 2013. DOI: 10.1016/j.jkss.2012.10.002 (cit. on p. 13).

[32] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997, pp. 1–19, ISBN: 9780070428072 (cit. on p. 13).

[33] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Computer Science*, vol. 2, no. 3, pp. 1–21, 2021. DOI: 10.1007/s42979-021-00592-x (cit. on pp. 13, 14, 17).

[34] A. Telikani, A. Tahmassebi, W. Banzhaf, and A. H. Gandomi, "Evolutionary machine learning: A survey", *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–35, 2021. DOI: 10.1145/3467477 (cit. on pp. 13, 16).

[35] M. A. El Mrabet, K. El Makkaoui, and A. Faize, "Supervised Machine Learning: A Survey", in *2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet)*, IEEE, 2021, pp. 1–10. DOI: 10.1109/CommNet52204.2021.9641998 (cit. on pp. 14, 19).

[36] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey", *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996. DOI: 10.1613/jair.301 (cit. on p. 14).

[37] S. Xu, B. Lu, M. Baldea, *et al.*, "Data cleaning in the process industries", *Reviews in Chemical Engineering*, vol. 31, no. 5, pp. 453–490, 2015. DOI: 10.1515/revce-2015-0022 (cit. on pp. 15, 30).

[38] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 2004, vol. 81, ISBN: 978-0471655749 (cit. on pp. 15, 23).

[39] K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, "2 - Data preprocessing", in *Computational Learning Approaches to Data Analytics in Biomedical Applications*, K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, Eds., Academic Press, 2020, pp. 7–27, ISBN: 978-0-12-814482-4. DOI: 10.1016/B978-0-12-814482-4.00002-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128144824000024 (cit. on pp. 15, 16).

[40] V. Kumar and S. Minz, "Feature selection: A literature review", *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014. DOI: 10.6029/smartcr.2014.03.007 (cit. on p. 15).

[41] K. Kunanbayev, I. Temirbek, and A. Zollanvari, "Complex Encoding", in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–6. DOI: 10.1109/IJCNN52387.2021.9534094 (cit. on p. 15).

[42] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria", *Annals of Mathematics and Artificial Intelligence*, vol. 41, no. 1, pp. 77–93, 2004. DOI: 10.1023/B:AMAI.0000018580.96245.c6 (cit. on p. 17).

[43] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm", *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 612–619, 2020 (cit. on p. 17).

[44] A. Mammone, M. Turchi, and N. Cristianini, "Support vector machines", *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 3, pp. 283–289, 2009. DOI: 10.1002/wics.49 (cit. on p. 17).

[45] E. García-Gonzalo, Z. Fernández-Muñiz, P. J. García Nieto, A. Bernardo Sánchez, and M. Menéndez Fernández, "Hard-Rock Stability Analysis for Span Design in Entry-Type Excavations with Learning Classifiers", *Materials*, vol. 9, no. 7, 2016, ISSN: 1996-1944. DOI: 10.3390/ma9070531. [Online]. Available: https://www.mdpi.com/1996-1944/9/7/531 (cit. on p. 17).

[46] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice", *Neurocomputing*, vol. 415, pp. 295–316, 2020. DOI: 10.1016/j.neucom.2020.07.061 (cit. on p. 18).

[47] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations", *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015. DOI: 10.5121/ijdkp.2015.5201 (cit. on pp. 19, 21).

[48] M. Grandini, E. Bagli, and G. Visani, *Metrics for Multi-Class Classification: an Overview*, 2020. DOI: 10.48550/ARXIV.2008.05756. [Online]. Available: https://arxiv.org/abs/2008.05756 (cit. on pp. 20, 21).

[49] T. Fawcett, "An introduction to ROC analysis", *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006. DOI: 10.1016/j.patrec.2005.10.010 (cit. on p. 21).

[50] V. Bewick, L. Cheek, and J. Ball, "Statistics review 13: Receiver operating characteristic curves", *Critical care*, vol. 8, no. 6, pp. 1–5, 2004. DOI: 10.1186/cc3000 (cit. on p. 22).

[51] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation", *Caspian journal of internal medicine*, vol. 4, no. 2, p. 627, 2013 (cit. on p. 22).

[52] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019, vol. 793, ISBN: 978-0-470-52679-8 (cit. on pp. 22–27, 30, 37, 45).

[53] R. Houari, A. Bounceur, A. K. Tari, and M. T. Kecha, "Handling Missing Data Problems with Sampling Methods", in *2014 International conference on advanced networking distributed systems and applications*, IEEE, 2014, pp. 99–104. DOI: 10.1109/INDS.2014.25 (cit. on pp. 22, 27, 28).

[54] W. R. Myers, "Handling Missing Data in Clinical Trials: An Overview", *Drug information journal: DIJ/Drug Information Association*, vol. 34, no. 2, pp. 525–533, 2000. DOI: 10.1177/009286150003400221 (cit. on p. 22).

[55] B. J. Wells, K. M. Chagin, A. S. Nowacki, and M. W. Kattan, "Strategies for Handling Missing Data in Electronic Health Record Derived Data", *Egems*, vol. 1, no. 3, 2013. DOI: 10.13063/2327-9214.1035 (cit. on pp. 22, 38).

[56] M. L. Bell and D. L. Fairclough, "Practical and statistical issues in missing data for longitudinal patient-reported outcomes", *Statistical methods in medical research*, vol. 23, no. 5, pp. 440–459, 2014. DOI: 10.1177/0962280213476378 (cit. on p. 22).

[57] D. B. Rubin, "Inference and Missing Data", *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976. DOI: 10.2307/2335739 (cit. on pp. 23, 32).

[58] C. K. Enders, *Applied Missing Data Analysis*. Guilford Publications, 2022, ISBN: 9781462549863 (cit. on pp. 24–30, 70).

[59] S. Fielding, P. M. Fayers, and C. R. Ramsay, "Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches", *Health and Quality of Life Outcomes*, vol. 7, no. 1, pp. 1–10, 2009. DOI: 10.1186/1477-7525-7-57 (cit. on p. 24).

[60] C. Xiong, K. Zhu, K. Yu, and J. P. Miller, "8 - Statistical Modeling in Biomedical Research: Longitudinal Data Analysis", in *Essential Statistical Methods for Medical Statistics*, C. Rao, J. Miller, and D. Rao, Eds., Boston: North-Holland, 2011, pp. 235–268, ISBN: 978-0-444-53737-9. DOI: 10.1016/B978-0-444-53737-9.50011-6. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444537379500116 (cit. on p. 25).

[61] C. Mack, Z. Su, and D. Westreich, *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition*, 2018. [Online]. Available: http://europepmc.org/books/NBK493611 (cit. on p. 26).

[62] R. J. Little, "A Test of Missing Completely at Random for Multivariate Data with Missing Values", *Journal of the American statistical Association*, vol. 83, no. 404, pp. 1198–1202, 1988. DOI: 10.1080/01621459.1988.10478722 (cit. on p. 26).

[63] H. Y. Chen and R. Little, "A test of missing completely at random for generalised estimating equations with missing data", *Biometrika*, vol. 86, no. 1, pp. 1–13, 1999, ISSN: 00063444 (cit. on p. 26).

[64] K. H. Kim and P. M. Bentler, "Tests of homogeneity of means and covariance matrices for multivariate incomplete data", *Psychometrika*, vol. 67, no. 4, pp. 609–623, 2002. DOI: 10.1007/BF02295134 (cit. on p. 26).

[65] Y. Dong and C.-Y. J. Peng, "Principled missing data methods for researchers", *SpringerPlus*, vol. 2, no. 1, pp. 1–17, 2013. DOI: 10.1186/2193-1801-2-222 (cit. on p. 26).

[66] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice", *Statistics in medicine*, vol. 30, no. 4, pp. 377–399, 2011. DOI: 10.1002/sim.4067 (cit. on p. 27).

[67] J. W. Graham *et al.*, "Missing Data Analysis: Making It Work in the Real World", *Annual review of psychology*, vol. 60, no. 1, pp. 549–576, 2009. DOI: 10.1146/annurev.psych.58.110405.085530 (cit. on p. 27).

[68] J. L. Peugh and C. K. Enders, "Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement", *Review of educational research*, vol. 74, no. 4, pp. 525–556, 2004. DOI: 10.3102/00346543074004525 (cit. on pp. 27–30).

[69] P. Sentas and L. Angelis, "Categorical missing data imputation for software cost estimation by multinomial logistic regression", *Journal of Systems and Software*, vol. 79, no. 3, pp. 404–414, 2006. DOI: 10.1016/j.jss.2005.02.026 (cit. on p. 28).

[70] R. R. Andridge and R. J. Little, "A Review of Hot Deck Imputation for Survey Non-response", *International statistical review*, vol. 78, no. 1, pp. 40–64, 2010. DOI: 10.1111/j.1751-5823.2010.00103.x (cit. on p. 28).

[71] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning", *Journal of Big Data*, vol. 8, no. 1, pp. 1–37, 2021. DOI: 10.1186/s40537-021-00516-9 (cit. on pp. 28, 29).

[72] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: A critical evaluation", *BMC medical informatics and decision making*, vol. 16, no. 3, pp. 197–208, 2016. DOI: 10.1186/s12911-016-0318-z (cit. on p. 29).

[73] A. Aleryani, W. Wang, and B. De La Iglesia, "Multiple Imputation Ensembles (MIE) for Dealing with Missing Data", *SN Computer Science*, vol. 1, no. 3, pp. 1–20, 2020. DOI: 10.1007/s42979-020-00131-0 (cit. on pp. 29, 34).

[74] S. Van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R", *Journal of statistical software*, vol. 45, pp. 1–67, 2011. DOI: 10.18637/jss.v045.i03 (cit. on pp. 29, 54, 55, 82, 130).

[75] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?", *International journal of methods in psychiatric research*, vol. 20, no. 1, pp. 40–49, 2011. DOI: 10.1002/mpr.329 (cit. on p. 29).

[76] Y. Zhao and Q. Long, "Multiple imputation in the presence of high-dimensional data", *Statistical Methods in Medical Research*, vol. 25, no. 5, pp. 2021–2035, 2016. DOI: 10.1177/0962280213511027 (cit. on p. 29).

[77] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature", *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014. DOI: 10.5194/gmd-7-1247-2014,2014 (cit. on p. 30).

[78] R. Razavi-Far, B. Cheng, M. Saif, and M. Ahmadi, "Similarity-learning information-fusion schemes for missing data imputation", *Knowledge-Based Systems*, vol. 187, p. 104 805, 2020. DOI: 10.1016/j.knosys.2019.06.013 (cit. on pp. 31, 33, 35, 37, 39).

[79] P. Mishra, K. D. Mani, P. Johri, and D. Arya, "FCMI: Feature Correlation based Missing Data Imputation", *arXiv preprint arXiv:2107.00100*, 2021. DOI: 10.48550/ARXIV.2107.00100 (cit. on pp. 32, 35, 39).

[80] T. Siswantining, S. M. Soemartojo, D. Sarwinda, *et al.*, "Application of Sequential Regression Multivariate Imputation Method on Multivariate Normal Missing Data", in *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, IEEE, 2019, pp. 1–6. DOI: 10.1109/ICICoS48119.2019.8982423 (cit. on p. 32).

[81] S. Faisal and G. Tutz, "Nearest neighbor imputation for categorical data by weighting of attributes", *Information Sciences*, vol. 592, pp. 306–319, 2022. DOI: 10.1016/j.ins.2022.01.056 (cit. on pp. 33, 36, 39).

[82] R. K. Bania and A. Halder, "R-Ensembler: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data", *Computer methods and programs in biomedicine*, vol. 184, p. 105 122, 2020. DOI: 10.1016/j.cmpb.2019.105122 (cit. on pp. 33, 39).

[83] P. Keerin and T. Boongoen, "Improved KNN Imputation for Missing Values in Gene Expression Data", *Computers, Materials & Continua*, 2021. DOI: 10.32604/cmc.2022.020261 (cit. on p. 33).

[84] R. Pan, T. Yang, J. Cao, K. Lu, and Z. Zhang, "Missing data imputation by K nearest neighbours based on grey relational structure and mutual information", *Applied Intelligence*, vol. 43, no. 3, pp. 614–632, 2015. DOI: 10.1007/s10489-015-0666-x (cit. on p. 33).

[85] C.-F. Tsai, M.-L. Li, and W.-C. Lin, "A class center based approach for missing value imputation", *Knowledge-Based Systems*, vol. 151, pp. 124–135, 2018. DOI: 10.1016/j.knosys.2018.03.026 (cit. on pp. 33, 34, 36).

[86] H. V. Bhagat and M. Singh, "NMVI: A data-splitting based imputation technique for distinct types of missing data", *Chemometrics and Intelligent Laboratory Systems*, vol. 223, p. 104 518, 2022. DOI: 10.1016/j.chemolab.2022.104518 (cit. on pp. 33, 34, 36, 54, 82).

[87] U. Yelipe, S. Porika, and M. Golla, "An efficient approach for imputation and classification of medical data values using class-based clustering of medical records", *Computers & Electrical Engineering*, vol. 66, pp. 487–504, 2018. DOI: 10.1016/j.compeleceng.2017.11.030 (cit. on pp. 33, 34).

[88] P. Sammulal, Y. Usha Rani, and A. Yepuri, "A class based clustering approach for imputation and mining of medical records (CBC-IM)", *IADIS International Journal on Computer Science & Information Systems*, vol. 12, no. 1, pp. 61–74, 2017, ISSN: 1646-3692 (cit. on pp. 33, 34).

[89] S.-d. Miao, S.-q. Li, X.-y. Zheng, *et al.*, "Missing Data Interpolation of Alzheimer's Disease Based on Column-by-Column Mixed Mode", *Complexity*, vol. 2021, 2021. DOI: 10.1155/2021/3541516 (cit. on pp. 34, 36, 39).

[90] S. I. Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique", *Journal of big Data*, vol. 7, no. 1, pp. 1–21, 2020. DOI: 10.1186/s40537-020-00313-w (cit. on pp. 34, 37).

[91] M. C. de Goeij, M. van Diepen, K. J. Jager, G. Tripepi, C. Zoccali, and F. W. Dekker, "Multiple imputation: Dealing with missing data", *Nephrology Dialysis Transplantation*, vol. 28, no. 10, pp. 2415–2420, 2013. DOI: 10.1093/ndt/gft221 (cit. on p. 34).

[92] P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, "Missing Data in Clinical Research: A Tutorial on Multiple Imputation", *Canadian Journal of Cardiology*, vol. 37, no. 9, pp. 1322–1331, 2021. DOI: 10.1016/j.cjca.2020.11.010 (cit. on pp. 34, 38).

[93] K. Blazek, A. van Zwieten, V. Saglimbene, and A. Teixeira-Pinto, "A practical guide to multiple imputation of missing data in nephrology", *Kidney International*, vol. 99, no. 1, pp. 68–74, 2021. DOI: 10.1016/j.kint.2020.07.035 (cit. on p. 34).

[94] H. Khan, X. Wang, and H. Liu, "Handling missing data through deep convolutional neural network", *Information Sciences*, vol. 595, pp. 278–293, 2022. DOI: 10.1016/j.ins.2022.02.051 (cit. on pp. 35, 37, 39).

[95] J. Yoon, J. Jordon, and M. Schaar, "GAIN: Missing Data Imputation using Generative Adversarial Nets", in *International conference on machine learning*, PMLR, 2018, pp. 5689–5698. DOI: 10.48550/ARXIV.1806.02920 (cit. on p. 35).

[96] S. J. Choudhury and N. R. Pal, "Imputation of missing data with neural networks for classification", *Knowledge-Based Systems*, vol. 182, p. 104 838, 2019. DOI: 10.1016/j.knosys.2019.07.009 (cit. on p. 35).

[97] X. Lai, X. Wu, L. Zhang, W. Lu, and C. Zhong, "Imputations of missing values using a tracking-removed autoencoder trained with incomplete data", *Neurocomputing*, vol. 366, pp. 54–65, 2019. DOI: 10.1016/j.neucom.2019.07.066 (cit. on p. 35).

[98] T. Schneider, "Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values", *Journal of climate*, vol. 14, no. 5, pp. 853–871, 2001. DOI: 10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2 (cit. on p. 35).

[99] O. Delalleau, A. Courville, and Y. Bengio, *Efficient EM Training of Gaussian Mixtures with Missing Data*, 2018. DOI: 10.48550/ARXIV.1209.0521. arXiv: 1209.0521 [cs.LG] (cit. on p. 35).

[100] J. M. Jerez, I. Molina, P. J. García-Laencina, *et al.*, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem", *Artificial intelligence in medicine*, vol. 50, no. 2, pp. 105–115, 2010. DOI: 10.1016/j.artmed.2010.05.002 (cit. on p. 37).

[101] M. M. Rahman and D. N. Davis, "Machine Learning-Based Missing Value Imputation Method for Clinical Datasets", in *IAENG transactions on engineering technologies*, Springer, 2013, pp. 245–257. DOI: 10.1007/978-94-007-6190-2_19 (cit. on p. 38).

[102] G. Madhu, B. Bharadwaj, G. Nagachandrika, and K. Vardhan, "A Novel Algorithm for Missing Data Imputation on Machine Learning", in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2019, pp. 173–177. DOI: 10.1109/ICSSIT46314.2019.8987895 (cit. on pp. 38, 39).

[103] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction", in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 202–208. DOI: 10.1109/ACII.2017.8273601 (cit. on pp. 38, 39).

[104] Z. Zhang, H. Fang, and H. Wang, "Multiple Imputation based Clustering Validation (MIV) for Big Longitudinal Trial Data with Missing Values in eHealth", *Journal of medical systems*, vol. 40, no. 6, pp. 1–9, 2016. DOI: 10.1007/s10916-016-0499-0 (cit. on pp. 38, 39).

[105] S. Tabarestani, M. Aghili, M. Eslami, *et al.*, "A distributed multitask multimodal approach for the prediction of Alzheimer's disease in a longitudinal study", *NeuroImage*, vol. 206, p. 116 317, 2020, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2019.116317. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811919309085 (cit. on pp. 38, 39).

[106] J. Yoon, W. R. Zame, and M. van der Schaar, "Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks", *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1477–1490, 2019. DOI: 10.1109/TBME.2018.2874712 (cit. on pp. 38, 39).

[107] J. R. Carpenter and M. Smuk, "Missing data: A statistical framework for practice", *Biometrical Journal*, vol. 63, no. 5, pp. 915–947, 2021. DOI: 10.1002/bimj.202000196 (cit. on p. 38).

[108] R. H. Groenwold, "Informative missingness in electronic health record systems: the curse of knowing", *Diagnostic and prognostic research*, vol. 4, no. 1, pp. 1–6, 2020. DOI: 10.1186/s41512-020-00077-0 (cit. on p. 38).

[109] G. Ras, N. Xie, M. van Gerven, and D. Doran, *Explainable Deep Learning: A Field Guide for the Uninitiated*, 2020. DOI: 10.48550/ARXIV.2004.14545. [Online]. Available: https://arxiv.org/abs/2004.14545 (cit. on p. 39).

[110] "The three ghosts of medical AI: Can the black-box present deliver?", *Artificial Intelligence in Medicine*, vol. 124, p. 102 158, 2022, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2021.102158 (cit. on p. 39).

[111] D. Dua and C. Graff, *UCI Machine Learning Repository*, 2017. [Online]. Available: http://archive.ics.uci.edu/ml (cit. on p. 40).

[112] V. Torra, J. Domingo-Ferrer, J. M. Mateo-Sanz, and M. Ng, "Regression for ordinal variables without underlying continuous variables", *Information Sciences*, vol. 176, no. 4, pp. 465–474, 2006. DOI: 10.1016/j.ins.2005.07.007 (cit. on p. 40).

[113] National Health and Nutrition Examination Survey Data. "Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS)". (2022), [Online]. Available: https://www.cdc.gov/nchs/nhanes/index.htm. (accessed: 01.07.2022) (cit. on p. 44).

[114] M. A. Clynes, N. C. Harvey, E. M. Curtis, N. R. Fuggle, E. M. Dennison, and C. Cooper, "The epidemiology of osteoporosis", *British Medical Bulletin*, 2020. DOI: 10.1093/bmb/ldaa005 (cit. on p. 44).

[115] EACTS | European Association for Cardio-Thoracic Surgery. "The Quality Improvement Programme (QUIP)". (2022), [Online]. Available: https://www.eacts.org/quip/quality-improvement-programme/. (accessed: 20.08.2022) (cit. on p. 46).

[116] T. Rockel, *missMethods: Methods for Missing Data*, R package version 0.3.0, 2022. [Online]. Available: https://CRAN.R-project.org/package=missMethods (cit. on p. 50).

[117] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating synthetic missing data: A review by missing mechanism", *IEEE Access*, vol. 7, pp. 11 651–11 667, 2019. DOI: 10.1109/ACCESS.2019.2891360 (cit. on pp. 50, 51, 130).

[118] B. Twala, "An empirical comparison of techniques for handling incomplete data using decision trees", *Applied Artificial Intelligence*, vol. 23, no. 5, pp. 373–405, 2009. DOI: 10.1080/08839510902872223 (cit. on p. 50).

[119] U. Garciarena and R. Santana, "An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers", *Expert Systems with Applications*, vol. 89, pp. 52–65, 2017. DOI: 10.1016/j.eswa.2017.07.026 (cit. on p. 50).

[120] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011 (cit. on p. 55).

Appendix

Additional Content

This appendix contains two tables with additional content concerning the two working real-world datasets, which might be of interest within the scope of this dissertation. The first one includes a more thorough characterisation of the Osteoporosis Dataset. The second table describes the Cardiothoracic Surgery Dataset in a more detailed manner.

Table A.1: Extensive characterisation of the Osteoporosis Dataset.

Variable	Description	Variable Type	MR
Demographics			
RIDAGEYR	Age in years, at the time of the screening interview.	Metric	0.0%
RIDRETH3	Race-ethnicity.	Nominal (6 categories)	0.0%
DMDEDUC2	Education level, only measured in adults aged 20 and over.	Ordinal (5 categories)	0.0%
INDFMIN2	Annual family income, reported as a range value in dollars.	Ordinal (14 categories)	4.3%
Nutrition			
DRTVB6	Vitamin B6, in mg	Metric	7.5%
DRTVB12T	Vitamin B12, in mg	Metric	7.5%
DRTVC	Vitamin C, in mg	Metric	7.5%
DRTVD	Vitamin D (D2 + D3), in µg	Metric	7.9%
DRTVK	Vitamin K, in µg	Metric	7.5%
DRTCALC	Calcium, in mg	Metric	7.5%
Blood pressure			
BPXSY	Systolic Blood pressure, in mmHg	Metric	3.0%
BPXDI	Diastolic Blood pressure, in mmHg	Metric	3.5%
Anthropometrics			
BMXWT	Weight, in kg	Metric	0.5%
BMXHT	Height, in cm	Metric	0.7%
Physical fitness			
MGDCGSZ	Grip strength, in kg, obtained through the sum of the largest reading from each hand.	Metric	10.0%
Blood lipids			
LBDHDD	HDL-Cholesterol, in mg/dL	Metric	2.9%
LBDLDL	LDL-Cholesterol, in mg/dL	Metric	53.4%
LBXTC	Total Cholesterol, in mg/dL	Metric	2.9%
LBXTR	Triglycerides, in mg/dL	Metric	53.0%
Hormones			
LBXTST	Testosterone, total, in ng/dL	Metric	3.7%
LBXEST	Estradiol, in pg/mL	Metric	8.4%
LBXSHBG	Sex Hormone Binding Globulin, in nmol/L	Metric	14.7%
Biochemistry			
LBXSCA	Calcium, in mg/dL	Metric	3.8%

Continuation of Table A.1

Variable	Description	Variable Type	MR
LBXVIDMS	Vitamin D Serum (D2 + D3), in nmol/L	Metric	2.4%
Physical activity			
PAXMTSD	Total physical activity (MIMS[1]/day)	Metric	9.8%
PAXMINSB	Sedentary behaviour (MIMS <1), in minutes/day	Metric	10.0%
PAXMINLPA	Light physical activity (1<MIMS <10), in minutes/day	Metric	10.0%
PAXMINMPA	Moderate physical activity (10<MIMS <30), in minutes/day	Metric	10.0%
PAXMINVPA	Vigorous physical activity (30<MIMS <45), in minutes/day	Metric	10.0%
PAXMINVVPA	Very vigorous physical activity (>45 MIMS), in minutes/day	Metric	10.0%
PAXSUMSB	Summation of acceleration values within sedentary behaviour, in MIMS/day	Metric	10.0%
PAXSUMLPA	Summation of acceleration values within light physical activity, in MIMS/day	Metric	10.0%
PAXSUMMPA	Summation of acceleration values within moderate physical activity (MIMS/day)	Metric	10.0%
PAXSUMVPA	Summation of acceleration values within vigorous physical activity (MIMS/day)	Metric	10.0%
PAXSUMVVPA	Summation of acceleration values within very vigorous physical activity (MIMS/day)	Metric	10.0%
Lifestyle			
PAXSWMD	Sleeping time, in minutes/day	Metric	9.8%
PAXLXSD	Light exposure, in LUX/day	Metric	9.8%

[1]Monitor-Independent Movement Summary (MIMS) is a non-proprietary, open-source, device-independent universal summary metric.

Table A.2: Extensive characterisation of the Cardiothoracic Surgery Dataset.

Variable	Variable Type	MR
Hospitalisation		
Date of admission	Date	0.0%
Date of death	Date	94.5%
Date of discharge	Date	0.1%
Date of surgery	Date	0.0%
Cardiac history		
Angina	Ordinal (5 categories)	0.3%
Dyspnoea	Ordinal (4 categories)	0.3%
Number of previous MIs	Ordinal (4 categories)	0.3%
Most recent MI	Ordinal (6 categories)	79.2%
Congestive heart failure	Binary	0.4%
Previous interventions		
Previous PCI	Ordinal (4 categories)	0.4%
Date of last PCI	Date	92.2%
Previous cardica. vascular	Nominal (4 categories)	0.2%
Date of last cardiac surgery	Date	86.9%
Pre-operative risk factors		
Weight	Metric	1.3%
Height	Metric	1.3%
Smoking history	Ordinal (3 categories)	0.9%
Diabetes treatment	Ordinal (4 categories)	0.8%
Hypercholesterolaemia	Binary	0.7%
Hypertension	Binary	0.7%
Renal disease	Nominal (5 categories)	0.8%
Last pre-operative creatinin (mg/dL)	Metric	0.8%
Last pre-operative creatinin.1 (µmol/L)	Metric	0.8%
Chronic lung disease	Ordinal (3 categories)	0.7%
Extra-cardiac arteriopathy	Nominal (3 categories)	0.9%
Cerebrovascular disease	Nominal (7 categories)	0.9%
Carotid bruits	Binary	0.9%
Neurological dysfunction	Binary	0.7%
Pre-operative heart rhythm	Nominal (7 categories)	0.8%
Pre-operative haemodynamics & catheterisation		
Left- or right-heart catheter	Nominal (3 categories)	1.4%
Date-of-last catheterisation	Date	25.2%
Number of diseased corona	Ordinal (4 categories)	11.0%
Left main stem disease	Binary	11.0%
Ejection fraction category	Ordinal (3 categories)	1.4%
Ejection fraction value	Metric	84.7%
PA systolic	Metric	89.6%
AV gradient	Metric	69.2%

Continuation of Table A.2

Variable	Variable Type	MR
LVEDP	Metric	99.9%
Mean PAWP / LA	Metric	99.9%
Pre-operative status & support		
IV inotropes	Binary	0.5%
IV nitrates / heparin of any	Binary	0.4%
Ventilated	Binary	0.7%
Cardiogenic shock	Binary	0.8%
Operation		
Operative urgency	Ordinal (4 categories)	0.3%
Main reason for urgency	Nominal (7 categories)	94.1%
Number of previous heart operations	Ordinal (6 categories)	0.3%
Procedure groups	Nominal (7 categories)	0.0%
Other cardiac procedures	Nominal (24 categories)	0.1%
Other non-cardiac procedures	Nominal (5 categories)	0.1%
Other operation description	Text	95.3%
Coronary surgery		
Arterial distal coronary anastamoses	Metric	60.8%
Venous distal coronary anastamoses	Metric	61.2%
Arteries used as grafts	Nominal (9 categories)	62.3%
Valve surgery		
Number of valve procedure	Metric	13.3%
Valve site	Nominal (4 categories)	44.4%
Stenosis	Binary	44.4%
Insufficiency	Ordinal (5 categories)	44.5%
Explant type	Nominal (6 categories)	44.5%
Native valve pathology	Nominal (11 categories)	44.4%
Reason for repeat valve surgery	Nominal (8 categories)	44.6%
Valve procedure	Binary	44.4%
Implant type	Nominal (6 categories)	44.4%
Implant code	Code	45.4%
Implant size	Metric	45.7%
Valve status	Ordinal (4 categories)	44.5%
Segments of the aorta	Ordinal (5 categories)	91.9%
Aortic procedure	Nominal (12 categories)	92.0%
Cardiopulmonary bypass	Nominal (3 categories)	0.2%
Predominant form of myocardial protection	Binary	19.0%
Cardioplgia: Infusion mode	Binary	21.5%
Cardioplgia: Solution	Binary	30.9%
Cardioplgia: Temperature	Binary	21.4%
Cardioplgia: Timing	Binary	21.6%
Intra-aortic balloon pump u	Nominal (4 categories)	0.5%

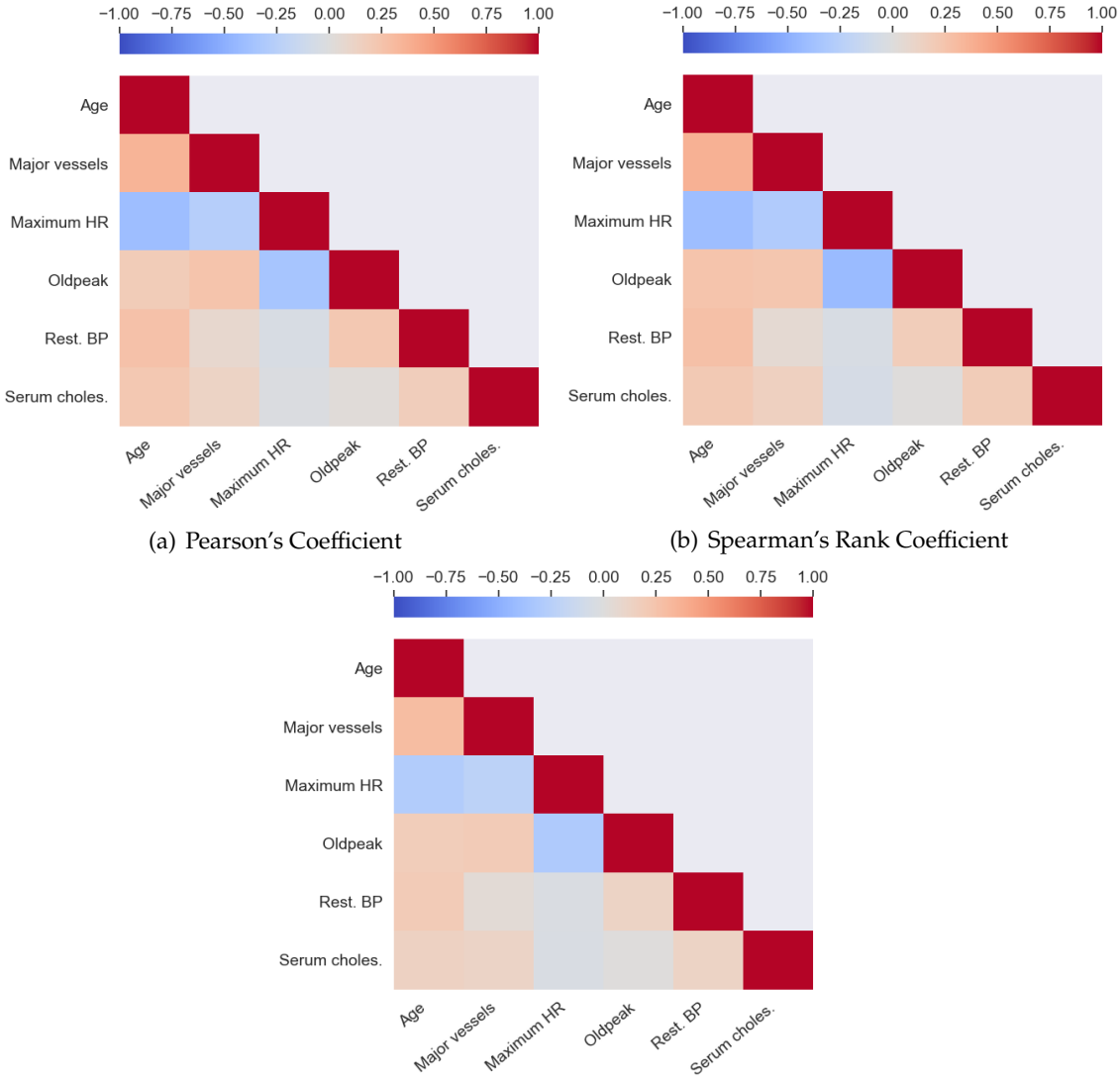
Continuation of Table A.2

Variable	Variable Type	MR
Non-cardioplegia myocardial protection	Nominal (5 categories)	97.5%
Reason for IABP use	Nominal (5 categories)	97.3%
Bypass time	Metric	1.5%
Total circulatory arrest time	Metric	4.5%
Cumulative X-clamp time	Metric	1.8%
Duration of surgery	Metric	0.5%
ICU stay	Metric	0.2%
Intermediate stay	Metric	2.6%
Cardiac Sugery Morbidity Scale		
CNS complications	Nominal (4 categories)	0.1%
Renal complications	Nominal (4 categories)	0.1%
Respiratory complications	Nominal (4 categories)	0.2%
Cardiac complications	Nominal (4 categories)	0.2%
Detailed cardiac complications	Nominal (9 categories)	66.4%
Bleeding complications	Nominal (4 categories)	0.1%
Infective complications	Nominal (4 categories)	0.2%
Other complications	Nominal (4 categories)	0.2%
Morbidity Index	Metric	0.5%
Discharge details		
Destination on discharge	Nominal (6 categories)	0.3%
Patient status at discharge	Binary	0.3%
Primary cause of death	Nominal (8 categories)	96.6%
Patient demographics and autocalculations		
Age	Metric	0.0%
Gender	Binary	0.0%
Hosp No	Code	0.0%
BMI	Metric	1.5%
BSA	Metric	1.4%
Additive EuroSCORE	Metric	0.0%
Euroscore results	Metric	0.9%
Logistic EuroSCORE	Metric	1.8%
Pre-operative stay	Metric	0.1%
Post-Operative stay	Metric	0.2%
Total hospital stay	Metric	0.2%
Status	Nominal (8 categories)	0.0%
Subsequent entry	Metric	95.5%

Appendix B

Complementary Results

This appendix includes twelve figures, each providing complementary information to the discussion of results held in Chapter 6. Furthermore, it contains a table detailing the values for the hyperparameters and parameters of the classifiers with the highest average AUROC and corresponding imputation methods, respectively.

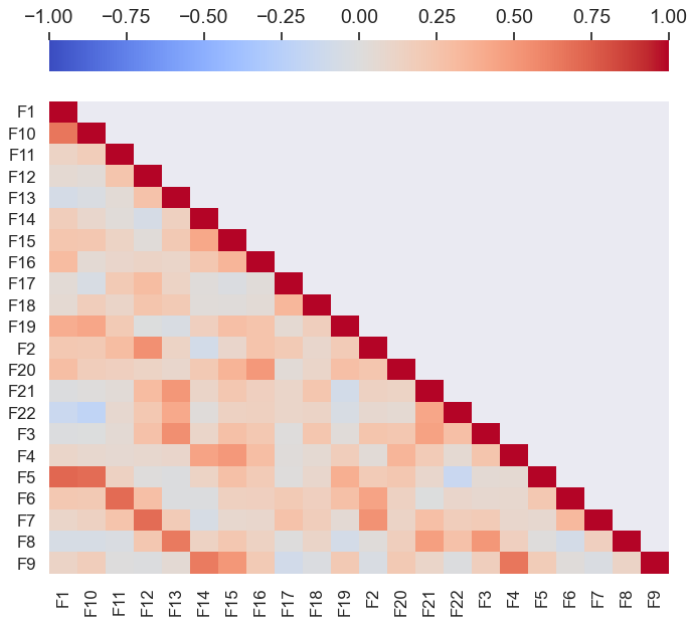


(a) Pearson's Coefficient

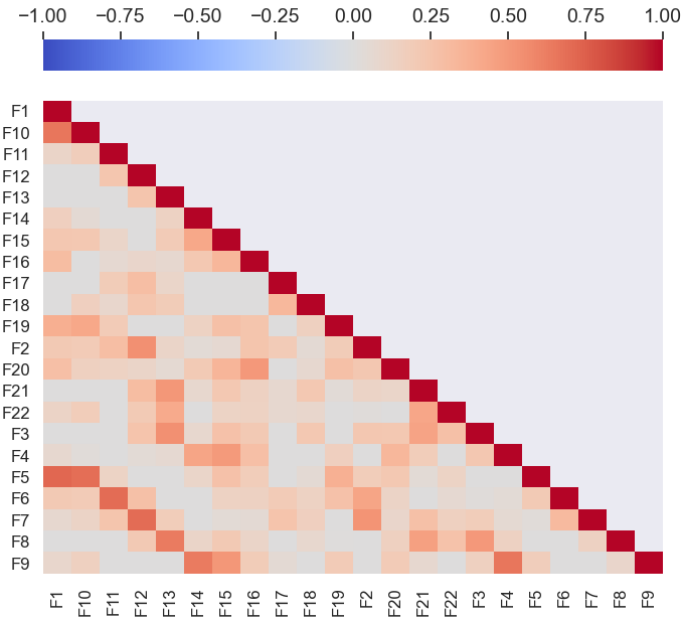
(b) Spearman's Rank Coefficient

(c) Kendall's Tau Coefficient (B)

Figure B.1: Correlation matrices of the numeric variables of the Statlog (Heart) Data Set obtain through different coefficients: (a) Pearson's coefficient, (b) Spearman's Rank coefficient, (c) Kendall's Tau coefficient.

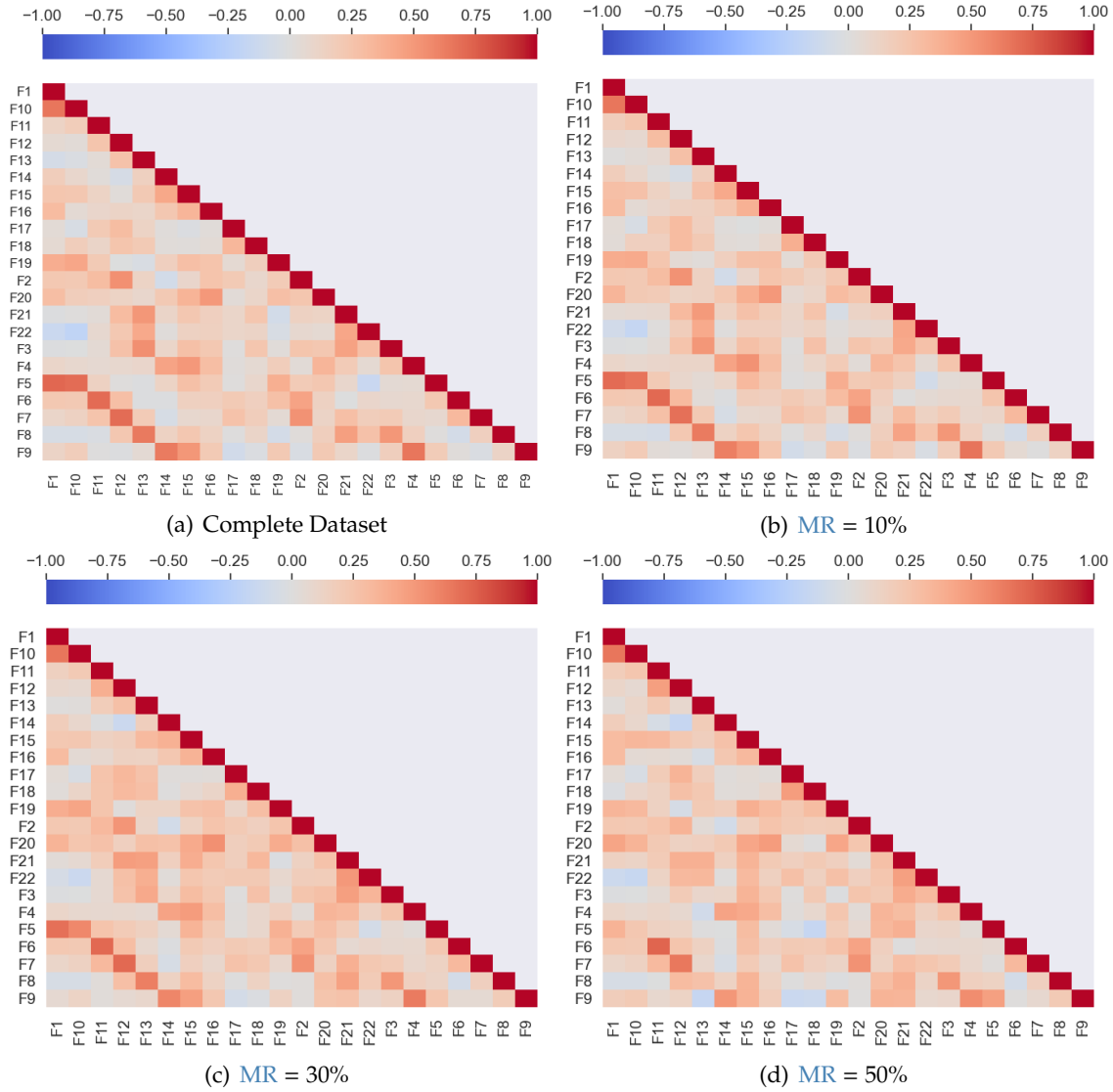


(a) Phi Coefficient



(b) Cramér's V (corrected)

Figure B.2: Correlation matrices of the SPECT Heart Data Set obtained through different coefficients: (a) Phi Coefficient, (b) Cramér's V (corrected).



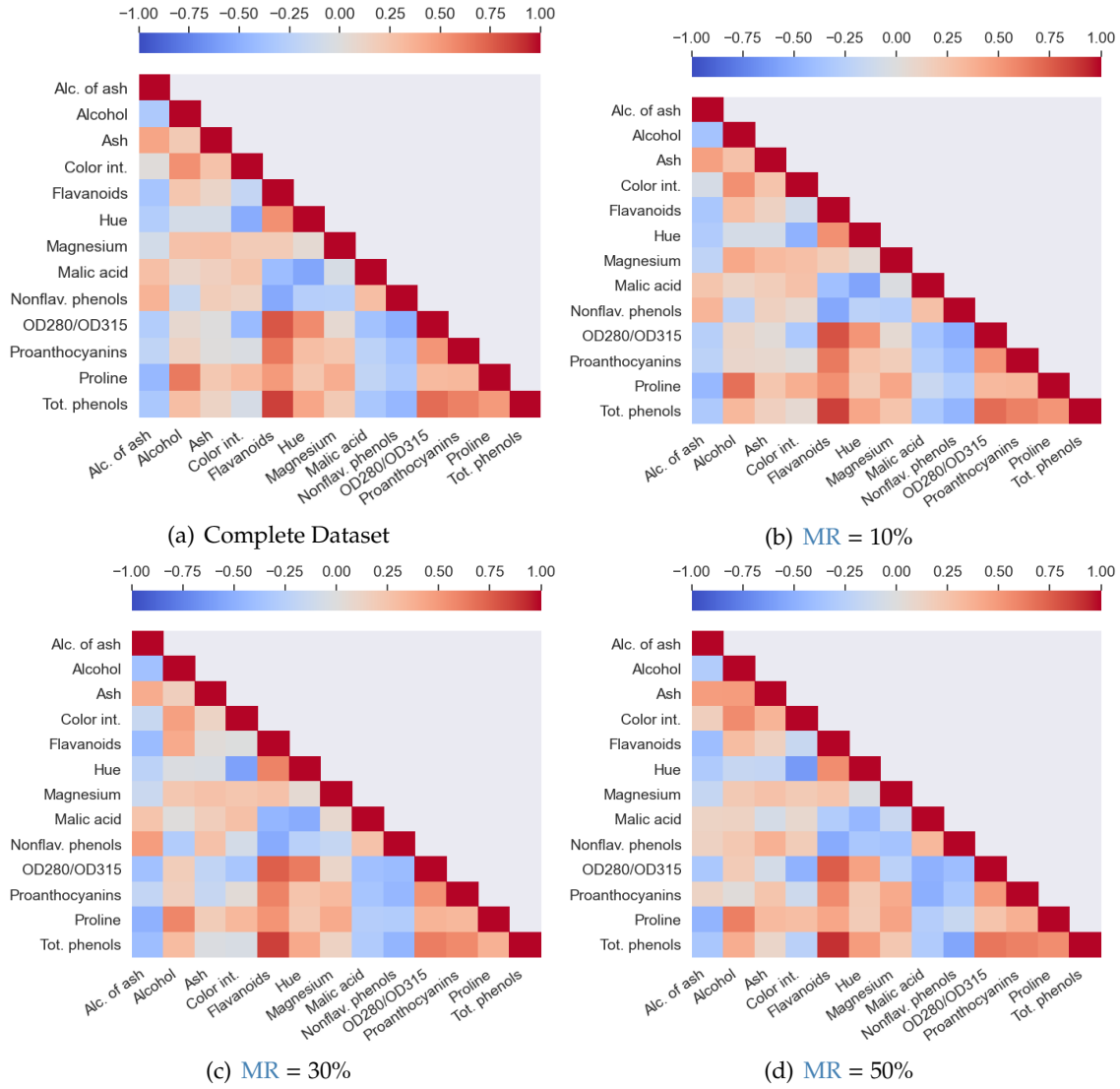
(a) Complete Dataset

(b) MR = 10%

(c) MR = 30%

(d) MR = 50%

Figure B.3: Correlation matrices of the SPECT Heart Data Set obtained for (a) the complete dataset and for different MRs under the MAR mechanism: (b) MR = 10%, (c) MR = 30%, (d) MR = 50%. The upper-triangle of the matrices was omitted for legibility purposes, as the information provided is redundant.



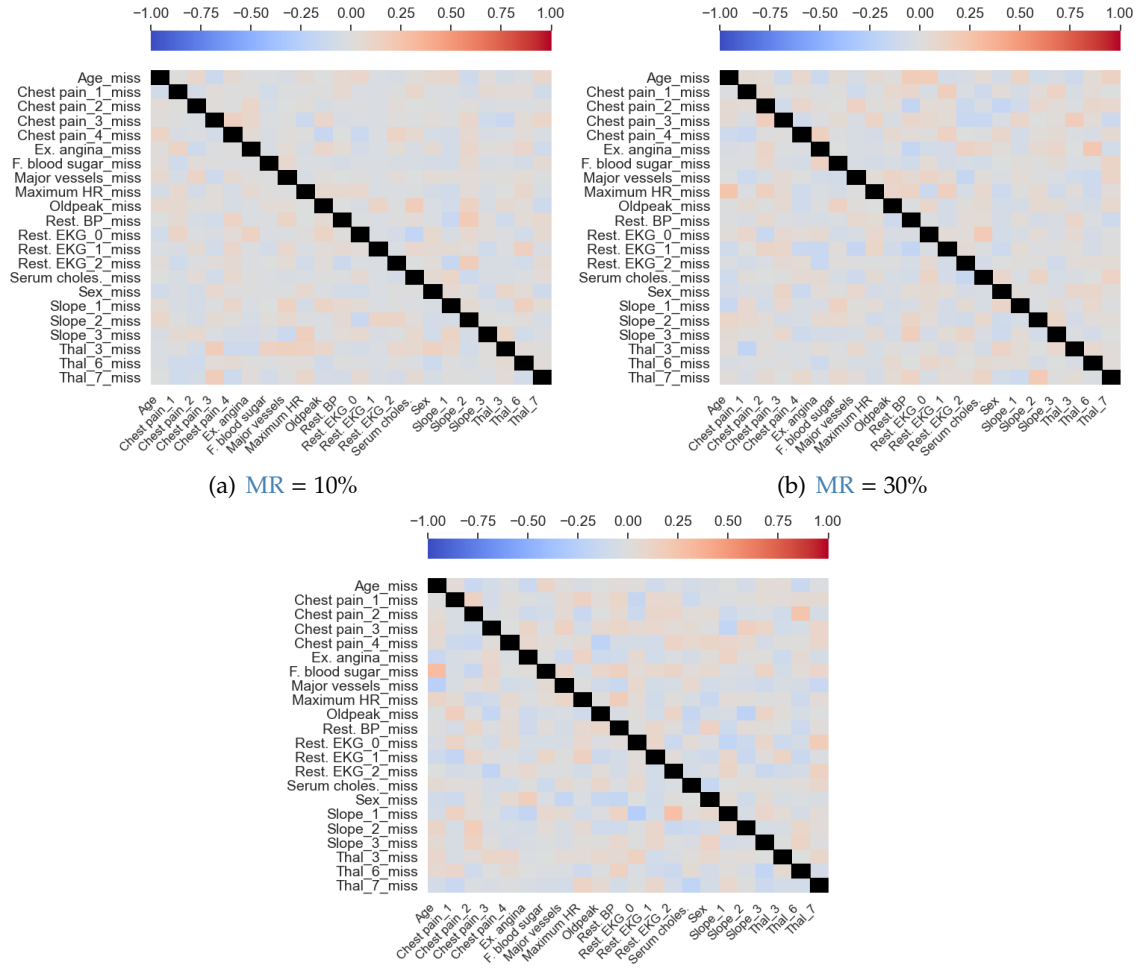
(a) Complete Dataset

(b) MR = 10%

(c) MR = 30%

(d) MR = 50%

Figure B.4: Correlation matrices of the Wine Data Set obtained for (a) the complete dataset and for different MRs under the MCAR mechanism: (b) MR = 10%, (c) MR = 30%, (d) MR = 50%. The upper-triangle of the matrices was omitted for legibility purposes, as the information provided is redundant.

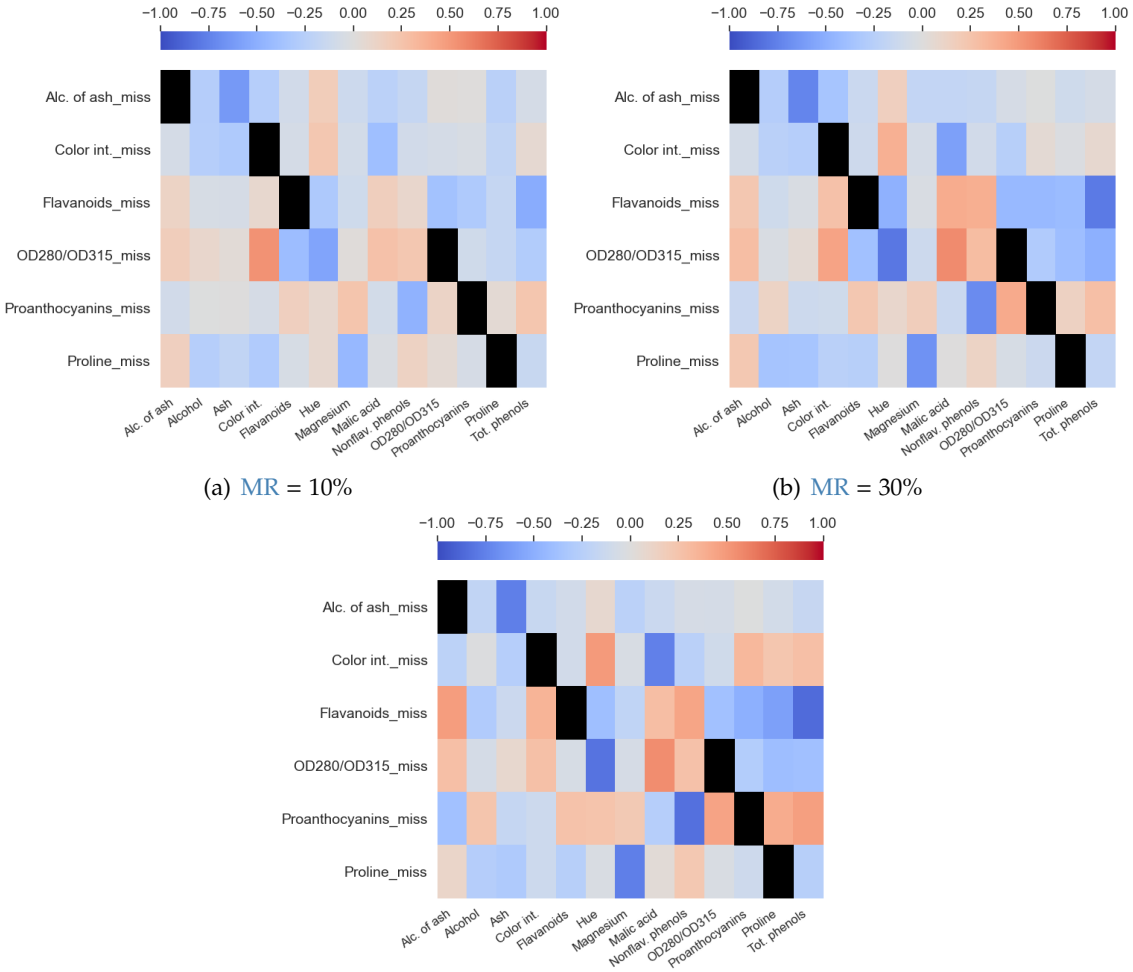


(a) MR = 10%

(b) MR = 30%

(c) MR = 50%

Figure B.5: Correlation matrices between the values and the missingness pattern of the Statlog (Heart) Data Set obtained for different MRs under the MCAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%. Missing features have the suffix "_miss". The black squares correspond to correlations that are impossible to compute.

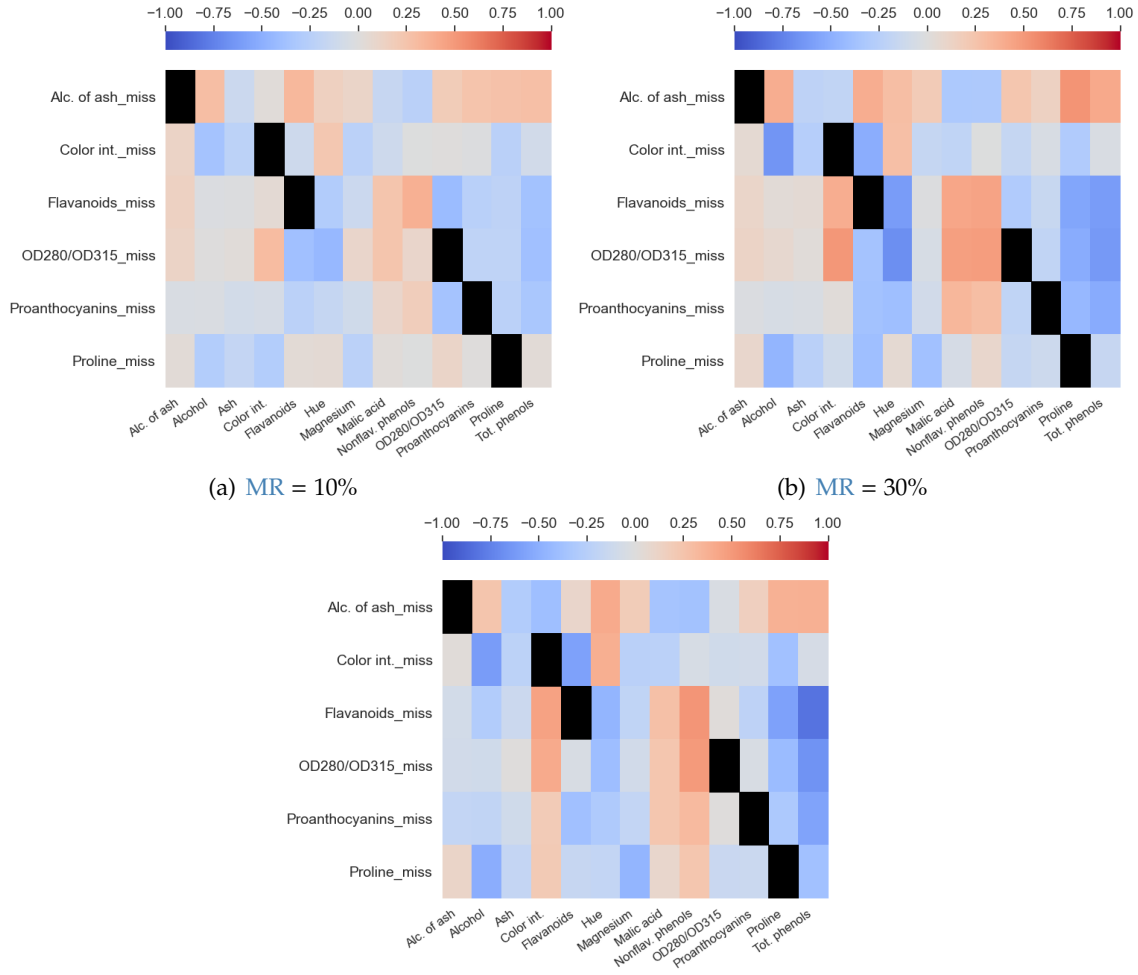


(a) MR = 10%

(b) MR = 30%

(c) MR = 50%

Figure B.6: Correlation matrices between the values and the missingness pattern of the Wine Data Set obtained for different MRs under the MAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%. Missing features have the suffix "_miss". The black squares correspond to correlations that are impossible to compute.

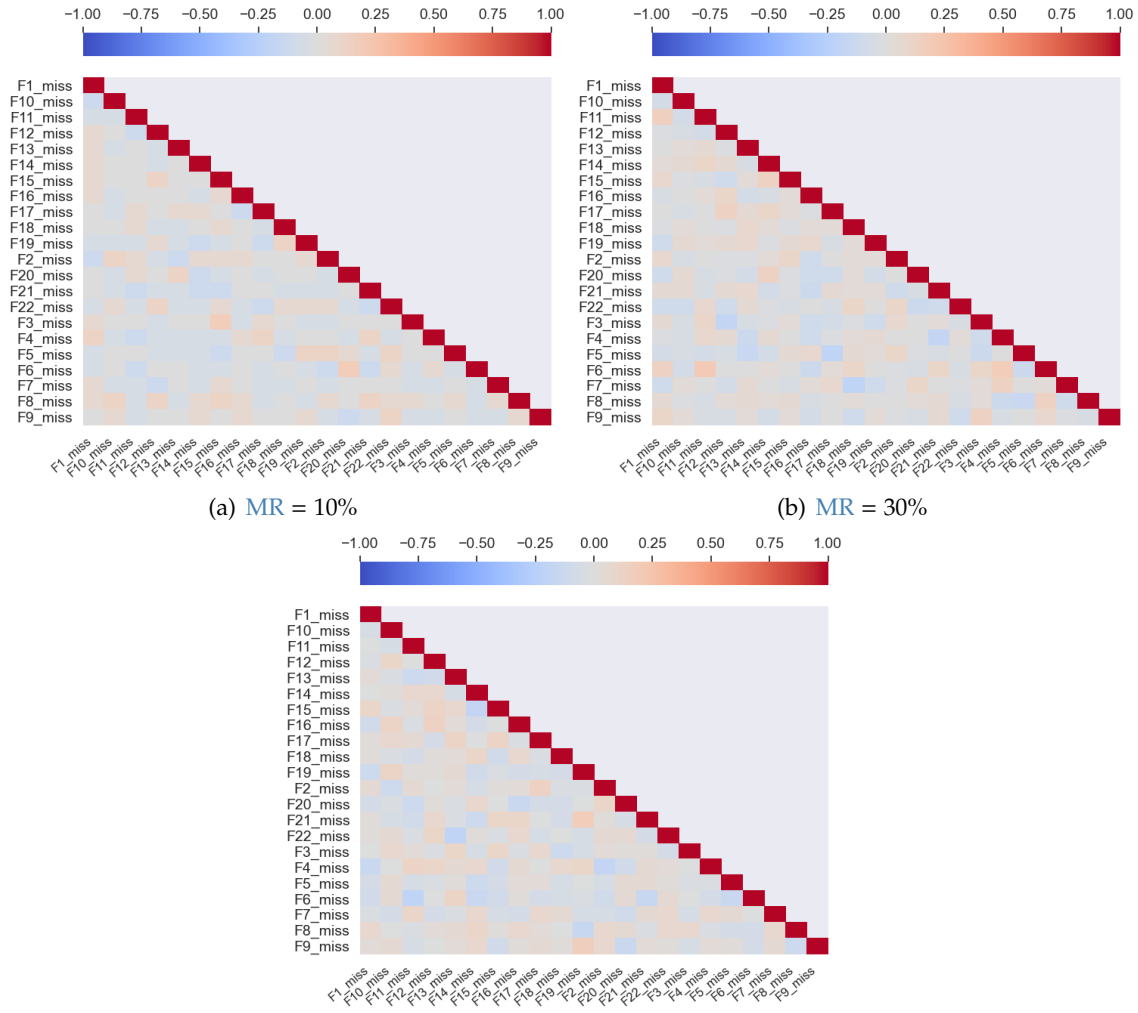


(a) MR = 10%

(b) MR = 30%

(c) MR = 50%

Figure B.7: Correlation matrices between the values and the missingness pattern of the Wine Data Set obtained for different MRs under the MNAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%. Missing features have the suffix "_miss". The black squares correspond to correlations that are impossible to compute.

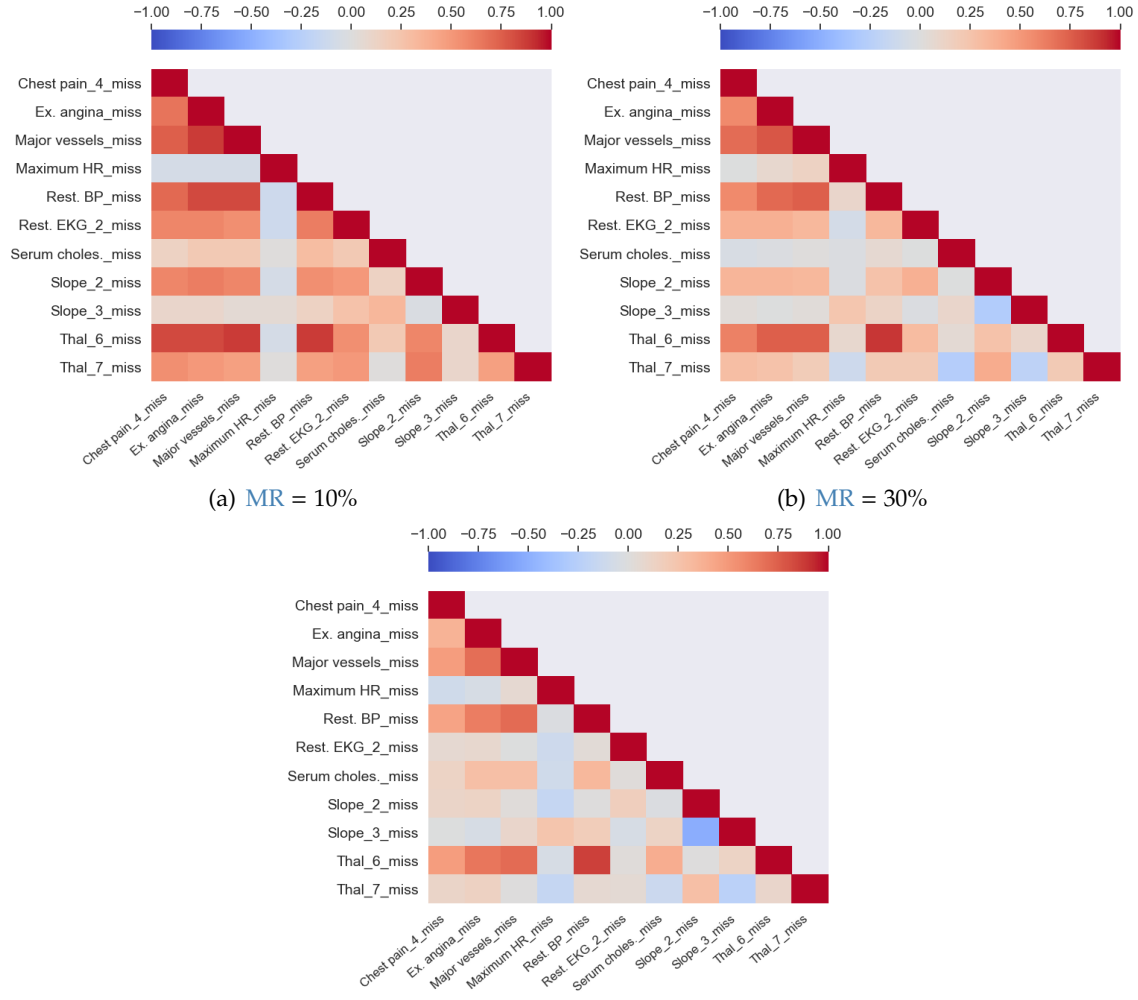


(a) MR = 10%

(b) MR = 30%

(c) MR = 50%

Figure B.8: Correlation matrices between the missingness patterns of the SPECT Heart Data Set obtained for different MRs under the MCAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%. The upper-triangle of the matrices was omitted for legibility purposes, as the information provided was redundant.

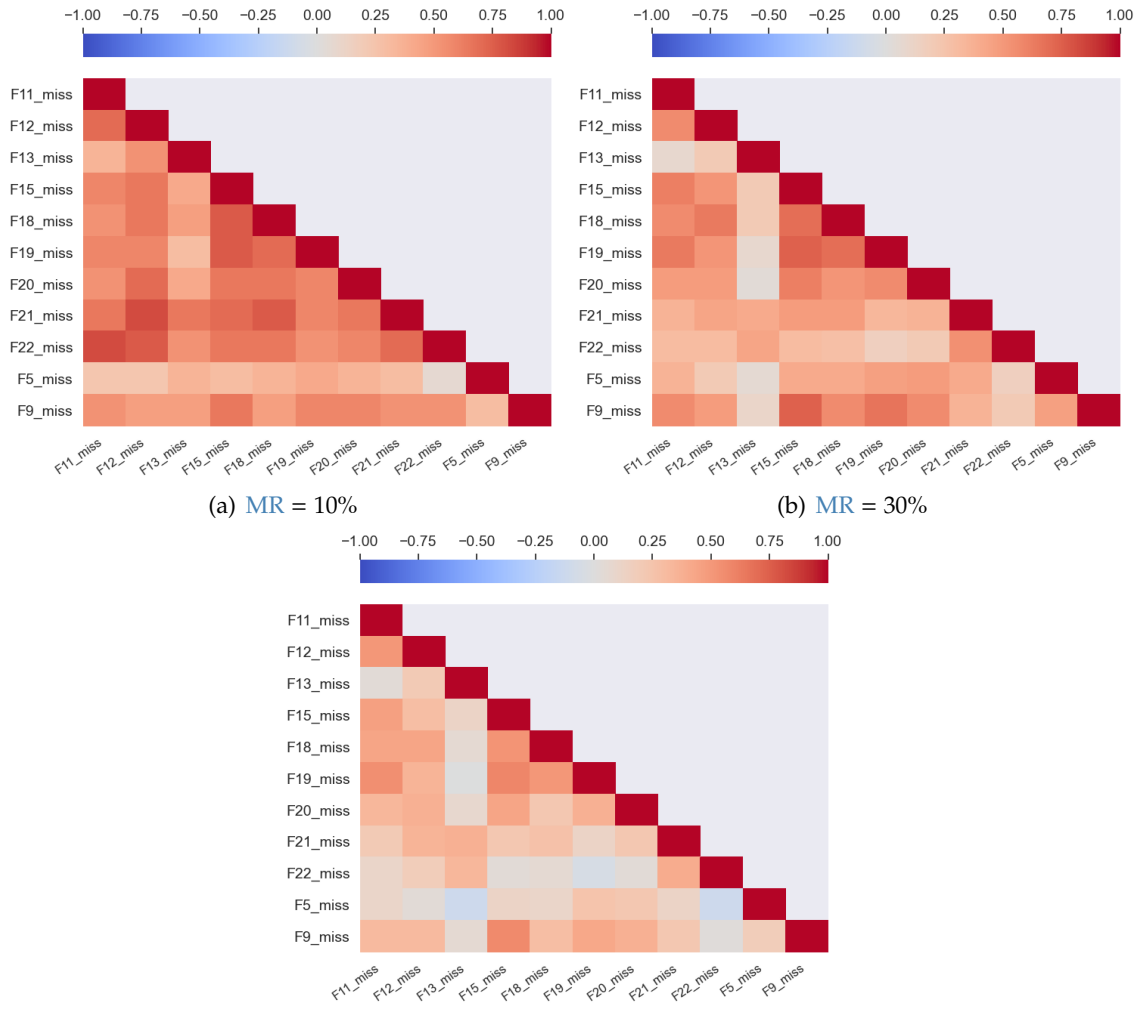


(a) MR = 10%

(b) MR = 30%

(c) MR = 50%

Figure B.9: Correlation matrices between the missingness patterns of the Statlog (Heart) Data Set obtained for different MRs under the MAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%. The upper-triangle of the matrices was omitted for legibility purposes, as the information provided was redundant.

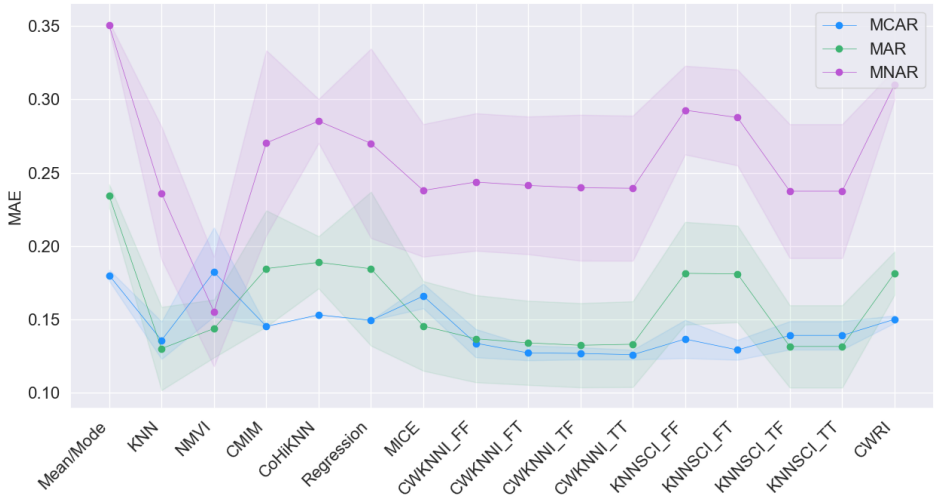


(a) MR = 10%

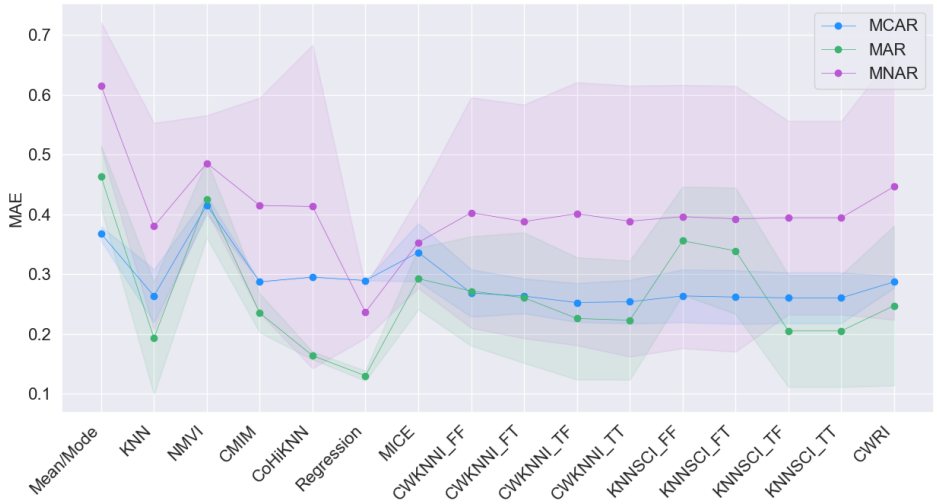
(b) MR = 30%

(c) MR = 50%

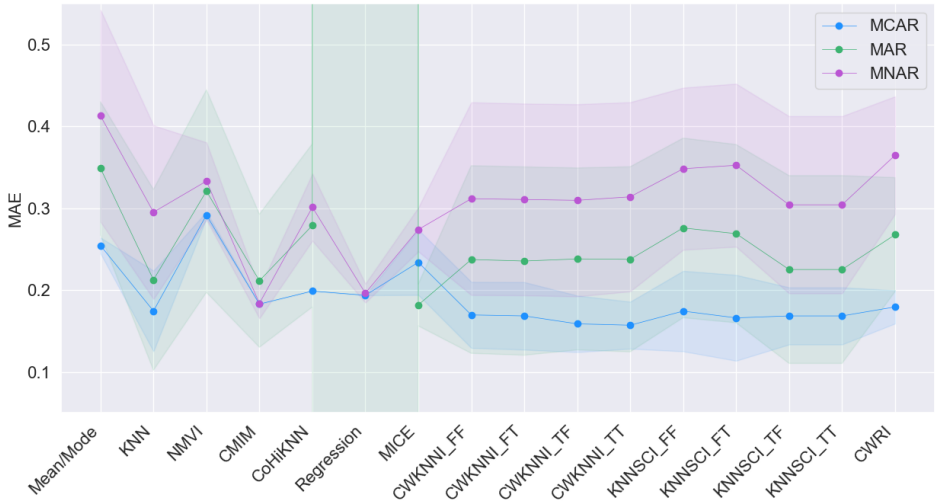
Figure B.10: Correlation matrices between the missingness patterns of the SPECT Heart Data Set obtained for different MRs under the MNAR mechanism: (a) MR = 10%, (b) MR = 30%, (c) MR = 50%. The upper-triangle of the matrices was omitted for legibility purposes, as the information provided is redundant.



(a) Wine Data Set

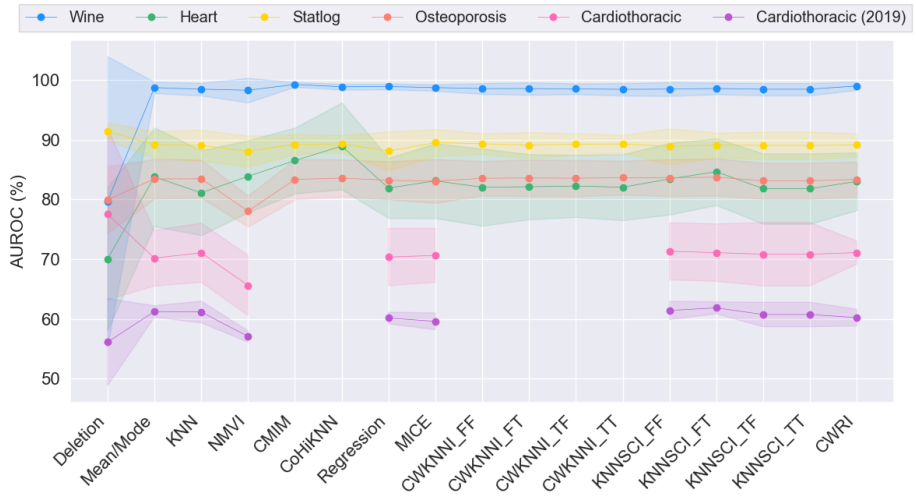


(b) SPECT Heart Data Set

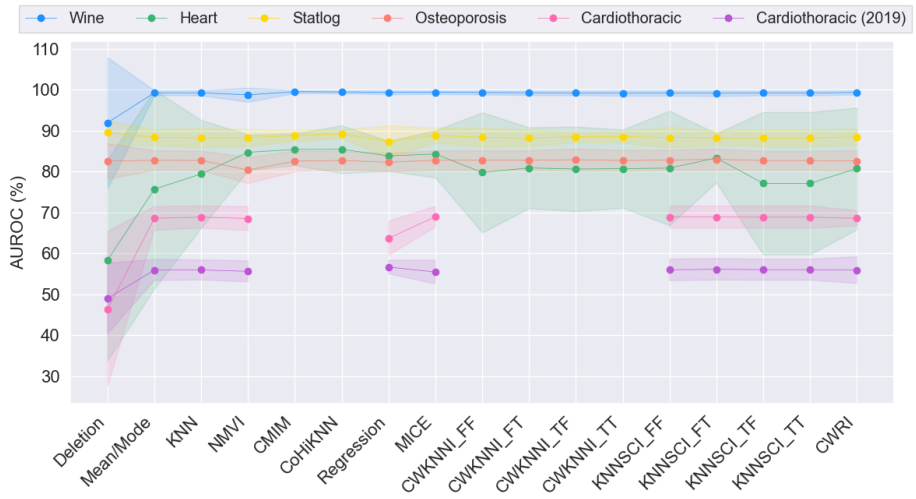


(c) Statlog (Heart) Data Set

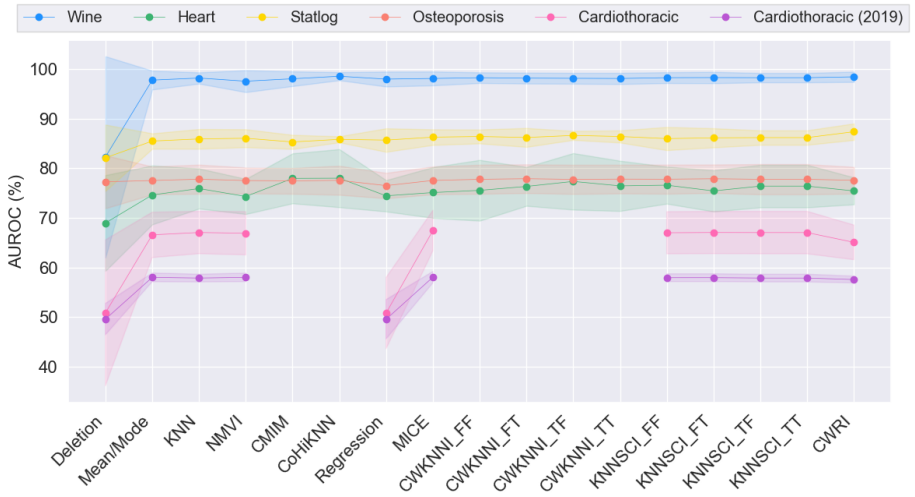
Figure B.11: Average MAE under all missingness mechanisms, for each synthetically generated dataset: (a) Wine Data Set, (b) SPECT Heart Data Set, (c) Statlog (Heart) Data Set.



(a) RF classifier



(b) SVM classifier



(c) NB classifier

Figure B.12: Average AUROC for every imputation method, using three diferent ML models: (a) RF classifier, (b) SVM classifier, (c) NB classifier.

Table B.1: Hyperparameters and parameters of the classifiers with the highest average AUROC and corresponding imputation methods, respectively.

Dataset	Classifier	Hyperparameters	Imputation Method	Parameters
Wine	RF	**random_state** = 42 **max_depth** = 3 **n_estimators** = 200 **criterion** = 'entropy' **min_samples_split** = 10 **min_samples_leaf** = 10	MICE	-
	SVM	**random_state** = 42 **class_weight** = 'balanced' **C** = 1 **gamma** = 1	CWRI	**percentage** = 0.8
	NB	**var_smoothing** = 1.0×10^{-4}	MICE	-
Heart	RF	**random_state** = 42 **max_depth** = 5 **n_estimators** = 100 **criterion** = 'entropy' **min_samples_split** = 20 **min_samples_leaf** = 5	KNNSCI_FT	**n_neighbors** = 10 **percentage** = 0.8 **initial_fill** = False **update** = True
	SVM	**random_state** = 42 **class_weight** = 'balanced' **C** = 0.1 **gamma** = 0.1	NMVI	-
	NB	**var_smoothing** = 1.0×10^{-4}	CWKNNI_TF	**n_neighbors** = 15 **percentage** = 0.2 **initial_fill** = True **update** = False
Statlog	RF	**random_state** = 42 **max_depth** = 5 **n_estimators** = 200 **criterion** = 'entropy' **min_samples_split** = 10 **min_samples_leaf** = 5	CWRI	**percentage** = 0.8
	SVM	**random_state** = 42 **class_weight** = 'balanced' **C** = 100 **gamma** = 0.01	CWRI	**percentage** = 0.8
	NB	**var_smoothing** = 1.0×10^{-4}	MICE	-

Continuation of Table B.1

Dataset	Classifier	Hyperparameters	Imputation Method	Parameters
Osteoporosis	RF	**random_state** = 42 **max_depth** = 9 **n_estimators** = 100 **criterion** = 'entropy' **min_samples_split** = 50 **min_samples_leaf** = 10	KNNSCI_FT	**n_neighbors** = 5 **percentage** = 0.6 **initial_fill** = False **update** = True
	SVM	**random_state** = 42 **class_weight** = 'balanced' **C** = 10 **gamma** = 0.1	KNNSCI_FT	**n_neighbors** = 10 **percentage** = 0.4 **initial_fill** = False **update** = True
	NB	**var_smoothing** = 1.0×10^{-4}	CWKNNI_FT	**n_neighbors** = 10 **percentage** = 0.2 **initial_fill** = False **update** = True
Cardiothoracic Surgery	RF	**random_state** = 42 **max_depth** = 6 **n_estimators** = 100 **criterion** = 'gini' **min_samples_split** = 20 **min_samples_leaf** = 25	KNNSCI_FF	**n_neighbors** = 15 **percentage** = 0.2 **initial_fill** = False **update** = False
	SVM	**random_state** = 42 **class_weight** = 'balanced' **C** = 0.1 **gamma** = 0.1	MICE	-
	NB	**var_smoothing** = 1.7×10^{-7}	MICE	-
Cardiothoracic Surgery (2019)	RF	**random_state** = 42 **max_depth** = 3 **n_estimators** = 100 **criterion** = 'gini' **min_samples_split** = 20 **min_samples_leaf** = 10	KNNSCI_FT	**n_neighbors** = 10 **percentage** = 0.6 **initial_fill** = False **update** = True
	SVM	**random_state** = 42 **class_weight** = 'balanced' **C** = 10 **gamma** = 1	Regression	-
	NB	**var_smoothing** = 1.7×10^{-7}	MICE	-

Appendix C

Python Libraries and R Packages

This appendix contains three tables. The first includes the Python libraries used throughout this dissertation and the second describes a selection of Python modules from the library scikit-learn. The third table presents the two R packages used.

Table C.1: Python libraries used throughout this dissertation.

Library	Version	Description
scikit-learn	0.24.1	Library for machine learning, providing functions for model fitting, data pre-processing, model selection, and model evaluation.
pandas	1.4.1	Library for data analysis and manipulation.
NumPy	1.22.1	Library for scientific computing, including functions to operate upon arrays.
SciPy	1.7.3	Library for scientific programming, including methods for statistical problems.
Matplotlib	3.3.4	Library for creating visualizations of data.
seaborn	0.11.1	Library for data visualization based on Matplotlib, with greater focus on the aesthetic.

Table C.2: Collection of the most relevant modules from the scikit-learn library within the scope of this work.

Module	Description
preprocessing.MinMaxScaler	Scales data to a given range (normalisation).
model_selection.StratifiedKFold	Splits the data into train and test set, preserving the proportion of instances for class each target.
model_selection.StratifiedGroupKFold	model_selection.StratifiedKFold variant with non-overlaping groups.
preprocessing.LabelEncoder	Ordinal encoding of categorical data.
preprocessing.OneHotEncoder	One-hot encoding of categorical data.
impute.SingleImputer	Univariate imputation with simple techniques.
model_selection.GridSearchCV	Performs hyperparameter tuning through a grid search strategy.
ensemble.RandomForestClassifier	RF classifier.
svm.SVC	SVM classifier.
naive_bayes.GaussianNB	Gaussian NB classifier.
metrics.roc_auc_score	Computes the AUROC from prediction scores.

Table C.3: R packages from the CRAN used in this dissertation.

Package	Description
missMethods	Methods for Missing Data: provides functions for the injection and handling of missing values, along with methods to evaluate missing data methods. Inspired in the work of Santos et al. [117].
mice	Multivariate Imputation by Chained Equations: implementation of the MICE algorithm as described by Van Buuren and Groothuis-Oudshoorn [74].

ANNEX I

Complementary Work

This annex contains the scientific paper "Addressing the Curse of Missing Data in Clinical Contexts: A Novel Approach to Correlation-Based Imputation", which presents the three novel correlation-based imputation methods developed during this work, as well as the results obtained in the comparative study. This paper was submitted to the *Journal of King Saud University - Computer and Information Sciences* and is currently under review.

Addressing the Curse of Missing Data in Clinical Contexts: A Novel Approach to Correlation-Based Imputation

Isabel Curioso[a,b], Ricardo Santos[a,b,*], Bruno Ribeiro[a], André Carreiro[a], Pedro Coelho[c,d], José Fragata[c,d] and Hugo Gamboa[a,b]

[a]Associação Fraunhofer Portugal Research, Rua Alfredo Allen 455/461, 4200-135 Porto, Portugal

[b]Laboratory for Instrumentation, Biomedical Engineering and Radiation Physics (LIBPhys-UNL), Physics Department, NOVA School of Science and Technology, 2829-516 Caparica, Portugal

[c]Comprehensive Health Research Center, NOVA Medical School, Campo Mártires da Pátria, 130, 1169-056 Lisboa, Portugal

[d]Hospital de Santa Marta, Centro Hospitalar Universitário Lisboa Central, Rua de Santa Marta, 50, 1169-023 Lisboa, Portugal

ARTICLE INFO

Keywords:
Missing Data
Missing Data Imputation
Correlation
Clinical Data
Machine Learning

ABSTRACT

Clinical data are essential in the medical domain. However, their heterogeneous nature leads to many data quality problems, notably missing values, which undermine the performance of Machine Learning-based clinical systems. Hence, there has been a growing interest in strategies that address this challenge in order to build trustworthy systems to improve the quality of care and benefit clinical decision-making. In particular, missing value imputation is a common approach. We propose three novel imputation techniques that leverage correlation in an innovative manner by exploring the relationship between values and missingness patterns. Experiments were carried out on three publicly available datasets, under three missingness mechanisms with different missing rates, and on two real-world medical datasets. The imputation precision and the classification performance of the proposed techniques were evaluated in a comprehensive comparative study, which included diverse existing methods. The developed techniques outperformed state-of-the-art methods on several assessments while overcoming current flaws shared by correlation-based imputation strategies in real-world medical problems.

1. Introduction

A growing and ageing population entails a broad amount of clinical information that needs to be organised and easily accessed by professionals. Since the analysis of paper-based data is time-consuming, digitalisation has emerged as a vital process towards optimising health care management, accompanied by the rise of valuable enablers such as Electronic Health Records (EHRs) (Ambinder, 2005).

An EHR contains thorough clinical information and can thus facilitate knowledge extraction. However, establishing relationships within the data to formulate a medical diagnosis still mostly relies on the physicians' experience. Artificial Intelligence (AI) is suitable for discovering patterns in vast datasets, an ability that could support and benefit clinical decision-making. Therefore, AI-based clinical systems will become an important instrument to assist professionals, leveraging all available information from the patient's journey.

Moreover, EHRs mirror the heterogeneous nature of clinical data, often collected through different procedures and stored in distinct formats. Unfortunately, with this variability also comes inconsistency. In fact, most real-world datasets are incomplete, which yields deleterious effects on Machine Learning (ML) models built thereon. In healthcare, reliable ML-based systems must be able to cope with missing values since their performance may influence clinical decision-making (Iranfar et al., 2021; Kang and Tian, 2018). This concern, along with the ubiquity of missing data in real-world databases, prompted a growing interest in developing strategies that address this challenge, particularly missing value imputation techniques.

In recent years, some authors have developed techniques that account for correlation when imputing missing values, stating that such a choice is beneficial. Even though correlation cannot assure a causal relationship between missingness and its source, it may still provide helpful insights for predicting missing data. The concept of correlation gains

*Corresponding author

isabel.curioso@fraunhofer.pt (I. Curioso); ricardo.santos@fraunhofer.pt (R. Santos); bruno.ribeiro@fraunhofer.pt (B. Ribeiro); andre.carreiro@fraunhofer.pt (A. Carreiro); pedro.coelho@chlc.min-saude.pt (P. Coelho); jose.fragata@nms.unl.pt (J. Fragata); hugo.gamboa@fraunhofer.pt (H. Gamboa)

ORCID(s): 0000-0002-4478-2476 (R. Santos)

relevance in clinical datasets, where distinct features are frequently different manifestations of the same physiological event or medical condition, consequently exhibiting a significant level of dependency. Although promising, the study of correlation within the context of missing value imputation is still scarce in biomedical research.

Therefore, this work exploits the correlation between attributes to address the challenges posed by missing data in real-world medical datasets, aiming to improve the robustness of ML-based systems. This paper contributes to the State of the Art with three novel correlation-based imputation techniques, which leverage not only the correlation between values but also the correlation between values and missingness patterns, an innovative and unique strategy. These techniques overcome the limitations of existing imputation methods in terms of their dependency on a complete data subset. Furthermore, a comprehensive comparative study was conducted to evaluate the effectiveness of the developed techniques.

The remaining of this paper is organised as follows. Related works, particularly state-of-the-art correlation-based imputation techniques, are briefly presented in Section 2. Section 3 covers materials and methods, including missingness mechanisms and correlation. Section 4 provides detailed descriptions of the proposed imputation techniques. Section 5 introduces the five datasets used throughout this work and the experimental setup. The obtained results are presented and discussed in Section 6. Lastly, Section 7 concludes this paper by reviewing its main findings along with perspectives for future work.

2. Related Work

There are two main approaches to address the challenges imposed by missing data, namely deletion and imputation.

The deletion methods, or techniques for ignoring missing data, are straightforward procedures based on completely recorded samples. As for imputation methods, their key purpose is to fill in, i.e. replace, the missing elements with predicted values, usually estimated from the observed data. These methods include a simple mean imputation, regression imputation and K-Nearest Neighbours (KNN) imputation (Little and Rubin, 2019).

As previously mentioned, several authors have recently turned their attention to correlation-based imputation techniques. Mishra et al. (2021) proposed an imputation method that replaces the missing elements in an attribute with predictions from regression models trained with features that are highly correlated with the incomplete attribute. Sefidian and Daneshpour (2020) also presented regression-based algorithms, called Correlation Maximisation-based Imputation Methods (CMIM), which attempted to maximise the correlation between the missing attributes and the remaining ones. Liu et al. (2019) developed the Correlation-based Hierarchical K-Nearest Neighbors (CoHiKNN) algorithm. This KNN-based algorithm utilises the correlation between attributes as weights to compute the distance between each incomplete record and all complete records. Khan et al. (2022) proposed the Convolutional Neural Network Imputation (CNNI) approach, which identifies existing correlations within a dataset to train a convolutional kernel that will replace all missing values.

Nevertheless, state-of-the-art imputation approaches still face limitations that ought to be overcome, such as the often unfeasible requirement for a complete subset, i.e. a subset without missing elements. Besides, a myriad of imputation methods is only validated on datasets with a controlled and synthetic missingness, which does not fully reflect the entropy of a real-world scenario.

3. Materials and Methods

3.1. Missingness Mechanisms

A nearly universal classification system for missing data problems was established by Rubin (1976), who pioneered the study on how the processes that cause missingness affect data analysis. Within this scope, Little and Rubin (2019) categorized the missingness mechanisms as Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing not at Random (MNAR). Although these mechanisms do not offer a causal explanation for the missingness, they describe generic relations between data and missing elements, which are important to understand when choosing the appropriate method to handle the missing values (Enders, 2022).

Let $X = (x_{ij})$ denote an $(N \times g)$ dataset without missing values, i.e. a complete dataset, where N is the number of samples or observations, and g is the number of features. x_{ij} is the value of variable X_j for observation i. If X contains missing data then the missingness indicator matrix $M = (m_{ij})$ is defined, such that $m_{ij} = 1$ if x_{ij} is missing and $m_{ij} = 0$ otherwise.

According to Little and Rubin (2019), the formal description of the missingness mechanisms relies on the conditional distribution of m_i given x_i, hereby represented as $f_{M|X}(m_i|x_i, \phi)$, where ϕ denotes unknown parameters.

In the MCAR mechanism, the missingness is completely unrelated to the data values, missing or observed. This mechanism verifies the equality

$$f_{M|X}(m_i|x_i, \phi) = f_{M|X}(m_i|x_i^*, \phi) \tag{1}$$

for all i and any distinct values (x_i, x_i^*) in the sample space of X. Accordingly, Equation 1 acknowledges that the probability of missingness on a variable X_j is not dependent on other measured variables nor on the values of X_j itself.

In the MAR mechanism, the missingness is related to the observed values of the data, but not the missing ones. Let $x_{(0)i}$ and $x_{(1)i}$ denote the observed and missing elements of x_i, respectively. This mechanism verifies the equality

$$f_{M|X}(m_i|x_{(0)i}, x_{(1)i}, \phi) = f_{M|X}(m_i|x_{(0)i}, x_{(1)i}^*, \phi) \tag{2}$$

for all i and any distinct values $(x_{(1)i}, x_{(1)i}^*)$ in the sample space of $X_{(1)}$. In conformity with Equation 2, the probability of missingness on a variable X_j depends solely on the values of another measured variable (or variables) but not on the values of X_j itself.

Finally, the MNAR mechanism's missingness is related to the unobserved data. According to Little and Rubin (2019), the distribution of m_i depends on the missing elements of x_i, i.e. Equation 2 does not apply for some sample i and some values $(x_{(1)i}, x_{(1)i}^*)$. This is the only mechanism that permits an association between the probability of missingness on a variable X_j and the values of X_j itself. Also, the missingness can depend on the observed values of the data as long as it still relates to the missing ones.

3.2. Correlation

Correlation is a measure of association between two variables, i.e. an indicator of how much a change in the magnitude of one variable is related to a change in the magnitude of another variable (Schober et al., 2018). Although knowing the values of a specific variable allows a better prediction of the values of a correlated variable, correlation does not necessarily assure causality.

Correlation coefficients are statistical measures of the degree of correlation between variables. There are several coefficients, each suitable for specific types of variables and with distinct underlying assumptions. Below, the coefficients used in this paper will be presented.

Pearson's (product-moment correlation) coefficient is one of the most used correlation measurements in medical research. Commonly denoted by r, this coefficient measures the strength of a linear relationship between two numeric, random variables X and Y, calculated by the following equation:

$$r = \frac{\text{cov}_{XY}}{\sigma_X \sigma_Y} \tag{3}$$

where cov_{XY} is the covariance value between X and Y, and σ_X and σ_Y are the standard deviations of variables X and Y, respectively.

Pearson's coefficient is a normalised covariance, scaled so that it ranges from -1 to +1, indicating a perfect negative and positive linear correlation, respectively. Pearson's correlation is only suitable for random numeric variables that follow a bivariate normal distribution (Schober et al., 2018; Akoglu, 2018).

The Phi coefficient, denoted by ϕ, is the equivalent of Pearson's coefficient, which measures the linear correlation between two binary variables X and Y. It can be calculated through Pearson's chi-square goodness-of-fit statistic for the 2×2 contingency table of the two variables:

$$\phi = \sqrt{\frac{\chi^2}{N}} \tag{4}$$

where χ^2 is the chi-square statistic and N is the number of observations. The Phi coefficient also ranges from -1 to +1.

The point biserial correlation coefficient, represented by r_{pbi}, measures the strength of association between a binary nominal variable Y and a numeric variable X:

$$r_{\text{pbi}} = \frac{\bar{X}_1 - \bar{X}_0}{\sigma_X} \sqrt{p_1 p_0} \tag{5}$$

where p_1 and $p_0 = 1 - p_1$ are respectively the proportion of samples with $Y = 1$ and $Y = 0$, $\bar{X}_1$ and $\bar{X}_0$ are respectively the means of X given $Y = 1$ and $Y = 0$, and σ_X is the standard deviation of X. This measurement also ranges from -1 to +1.

The chosen coefficients are all based on Pearson's correlation. Therefore, the correlation values obtained for relationships between different variable types are comparable.

4. Novel Correlation-Based Imputation Techniques

This paper proposes three novel techniques which leverage correlation when performing missing value imputation. These techniques aim to tackle some of the drawbacks found in the methods from the literature, such as the need for a complete subset and the evaluation on a single missingness mechanism. Furthermore, they exploit the concept of correlation, a promising but still not widely adopted approach in medical research. Rather than just resorting to the correlation between values, these methods investigate the potential benefits of considering the correlation between values and missingness patterns. If this correlation is strong, then the missing elements on the incomplete variable are confined to a particular segment in the distribution of the other variable's values. Such association may be helpful when computing estimates for the missing elements.

For this purpose, it is key to establish the procedure by which the matrices denoting the correlation between values and missingness patterns are obtained. Consider any dataset and let g denote the total number of features and $g_{\text{miss}} \leq g$ the number of features with missing values. g_j represents the jth feature, where $j \in \{1, 2, ..., g\}$, and $g_{\text{miss},i}$ represents the ith incomplete feature, where $i \in \{1, 2, ..., g_{\text{miss}}\}$. Furthermore, consider that C_{vm} is a $(g_{\text{miss}} \times g)$ matrix. For every pair $\{i, j\}$, with $g_{\text{miss},i} \neq g_j$, $C_{\text{vm}}[i, j]$ stores the correlation between the values of g_j and the missingness pattern of $g_{\text{miss},i}$, i.e. its binary missingness indicator.

Additionally, let C_{vv} denote the standard $(g \times g)$ correlation matrix. This matrix is computed through a pairwise deletion strategy, i.e. the correlation is calculated between the available values within each pair of attributes on an analysis-by-analysis basis.

4.1. Correlation Weighted K-Nearest Neighbour Imputation

The Correlation Weighted K-Nearest Neighbour Imputation (CWKNNI) method was inspired by the imputation technique CoHiKNN, proposed by Liu et al. (2019). However, instead of uniquely considering the correlation between values of different attributes, the weights in CWKNNI are obtained through a weighted average of the correlation between values and the correlation between values and missingness patterns. Additionally, CWKNNI overcomes the limitation of CoHiKNN in terms of its dependency on a complete data subset to impute the missing values, as it computes the correlation matrix through a pairwise deletion approach instead of performing listwise deletion.

The following step-by-step explanation provides the outline for CWKNNI:

1. Consider a dataset X, with g features and N instances.
2. Compute the $(g \times g)$ correlation matrix, denoted as C_{vv}, and the $(g_{\text{miss}} \times g)$ matrix, hereby denoted as C_{vm}, with the correlations between values and missingness patterns. This approach considers the absolute value of these correlations, thus accounting for the strength of the association, not its direction.
3. Order the features from lowest to highest missing rate. Imputation will be performed in this sequence, in a phased manner.
4. Let g_j denote the attribute being imputed. Create a subset $X_{\text{miss\_}j}$ consisting of samples where g_j is missing. Furthermore, create a subset $X_{\text{obs\_}j}$ with samples where g_j is observed. If $X_{\text{miss\_}j}$ is empty, skip to the next attribute.
5. For each instance in $X_{\text{miss\_}j}$ find the k nearest neighbours within $X_{\text{obs\_}j}$. A weighted euclidean distance which accounts for the presence of missing values is used as a distance measure:

$$d_{vt} = \sqrt{w_{\text{D}} \times \sum_{i \in O_g} (1 - w_{\text{C}i}) \times (v_i - t_i)^2} \tag{6}$$

where d_{vt} is the distance between samples v and t, v is an instance of $X_{\text{miss\_}j}$, t is a sample from $X_{\text{obs\_}j}$. v_i and t_i are the observed values from the ith attribute of v and t respectively. O_g is the set of indexes of the variables that are not missing in v nor in t. w_{D} is the quotient between the total number of features g and the dimension

of O_g. As for w_{Ci}, it is a weighted average of the correlation between g_j and g_i, and the correlation between the values of g_i and the missingness pattern of g_j:

$$w_{Ci} = p \times C_{vv}[i, j] + (1 - p) \times C_{vm}[i, j] \ , \qquad p \in [0, 1] \tag{7}$$

Note that $0 \leq w_{Ci} \leq 1$.

6. Replace the missing value of g_j in each instance of X_{miss_j} with a weighted prediction, in which the weights are the inverse of the computed distances d_{vl}: a closer neighbour has higher importance (i.e. weight) in the final prediction. The mean value is used for numeric variables, whereas the mode is used to impute binary attributes.
7. Repeat Steps 4-6 until all missing values from all features have been imputed.

The number of neighbours k and the percentage p, i.e. the weight placed on the correlation between values in comparison to the correlation between values and missingness pattern, are two parameters of CWKNNI.

This method also accepts a boolean parameter, **initial_fill**, which determines if an initial and temporary imputation with the standard KNN method is carried out. The values from this imputed dataset will be used for the calculation of both C_{vv} and C_{vm}. Furthermore, the subset X_{obs_j} will be formed by these KNN imputed samples. The number of samples within each X_{obs_j} remains the same; the only difference is that no instance has missing attributes. As for the sample being imputed, g_j will be its only missing feature.

Additionally, the boolean argument **update** controls if the imputed values will be used in subsequent phases of imputation instead of just considering the original values of each sample. Both matrices C_{vv} and C_{vm} are updated at the beginning of each phase, thus accounting for the newly imputed values. Note that the missingness patterns used to calculate C_{vm} do not change.

4.2. K-Nearest Neighbours Selected by Correlation Imputation

As with most KNN-based approaches, CWKNNI exhibits an increased computational cost for high dimensional data. The K-Nearest Neighbours Selected by Correlation Imputation (KNNSCI) method was created to overcome this drawback. For each variable to be imputed, this technique performs a pre-selection of features based on correlation, which reduces the dimensionality and facilitates its application on large datasets. The outline for this method is given below:

1-4. Equal to Steps 1-4 from CWKNNI.

5. Using the correlation matrices, compute the coefficient $r_{Cj,h}$ between attribute g_j and the hth attribute of X (apart from g_j). This coefficient is equal for all samples within the subset X_{miss_j} and is obtained through the following equation:

$$r_{Cj,h} = p \times C_{vv}[h, j] + (1 - p) \times C_{vm}[h, j] \ , \qquad p \in [0, 1] \tag{8}$$

Note that $0 \leq r_{Cj,h} \leq 1$.

6. Create a subset X_{obs_j} with samples where g_j is observed. Find the features where $r_{Cj,h} < 0.7$ and remove their columns from both X_{miss_j} and X_{obs_j}. The value of 0.7 was chosen based on the work of Schober et al. (2018), which stated that a coefficient above 0.7 indicates a strong correlation.
7. For each instance in X_{miss_j} find the k nearest neighbours within X_{obs_j} (after removing the columns with $r_{Cj,h} < 0.7$). A standard euclidean distance is used as a distance measure.
8. Replace the missing value on the attribute g_j in each instance of X_{miss_j} with a weighted prediction, in which the weights are the inverse of the computed euclidean distances: a closer neighbour has a higher importance in the final prediction. The mean value is used for numeric variables, whereas the mode is used to impute binary attributes.
9. Repeat Steps 4-8 until all missing values from all features have been imputed.

In addition to the number of neighbours k and the percentage p, KNNSCI also has the parameters **initial_fill** and **update**, which serve a similar purpose as in CWKNNI.

4.3. Correlation Weighted Regression Imputation

The Correlation Weighted Regression Imputation (CWRI) approach is a regression-based method in which the missing values are imputed with predictions drawn from distinct linear regression models. This technique finds several estimates for each missing value and combines them taking into account the correlational importance of the predictor variables. This correlational importance is obtained through a weighted average of the correlation between values and the correlation between values and missingness patterns.

A step-by-step outline of CWRI is presented:

1. Consider a dataset X, with g features and N instances. Additionally, let g_{miss} denote the number of attributes with missing values.
2. Compute the $(g \times g)$ correlation matrix, denoted as C_{vv}, and the $(g_{\text{miss}} \times g)$ matrix, whereby denoted as C_{vm}, with the correlations between values and missingness patterns. This approach leverages the absolute value of these correlations.
3. Build a $g \times g$ matrix, denoted as R, containing multiple linear regression models. Let g_i and g_j be two different attributes of X, with indexes i and j, respectively. For every pair $\{i, j\}$, where $i \neq j$, $R[i, j]$ stores a linear regression model in which g_i is the independent variable and g_j is the predictor.
4. Assemble the subset X_{inc} including all incomplete samples, i.e. instances with at least one missing value.
5. Consider any instance of X_{inc}, and let g_k denote its kth missing attribute. Using every non-missing feature g_t in this sample, obtain the corresponding predicted value for g_k through the regression model $R[k, t]$. If g_k is a numeric variable, the final imputation will be a weighted average of all predicted values. If g_k is binary, a weighted mode is applied. To obtain the weight given to each variable g_t, denoted as $w_{k,t}$, first calculate its correlational importance $w_{\text{C}k,t}$ through the following equation:

$$w_{\text{C}k,t} = p \times C_{\text{vv}}[t, k] + (1 - p) \times C_{\text{vm}}[t, k] \,, \qquad p \in [0, 1] \tag{9}$$

   Note that $0 \leq w_{\text{C}k,t} \leq 1$. After computing $w_{\text{C}k,t}$ for every non-missing feature, $w_{k,t}$ is a simple normalisation:

$$w_{k,t} = \frac{1}{\sum_t w_{\text{C}k,t}} \times w_{\text{C}k,t} \tag{10}$$

6. Repeat Step 5 until all incomplete samples have been imputed.

The percentage p is the only parameter of CWRI.

5. Experiments

5.1. Datasets

Three complete and publicly available datasets were selected from the UCI Machine Learning Repository: **Wine Data Set**, only comprising numeric variables; **SPECT Heart Data Set**, which solely includes binary attributes; **Statlog (Heart) Data Set**, a mixed-type dataset. The selection process was primarily based on the type of variables of each dataset, as it is essential to ensure that these methods perform a suitable and efficient imputation regardless of the attribute's type.

Within this paper, multiclass nominal variables were one-hot encoded, and ordinal encoding was applied to ordinal attributes. Hence, assessing the imputation precision is only relevant for numeric and binary variables.

Synthetic missing values were injected into these three datasets, under all three missingness mechanisms, with three different missing rates (10%, 30%, and 50%) defined for every attribute. In the MCAR mechanism, all features were incomplete and shared the same missing rate, whereas in the remaining two mechanisms, only 50% of the features were incomplete but also had the same missing rate. A total of nine synthetic datasets were generated per UCI Machine Learning Repository dataset. For this end, the R package missMethods (Rockel, 2022) was used, as it supplies functions for injecting missing data. This manipulation enabled a comprehensive study that included different missing rates and missingness mechanisms.

Then, since the main focus of this paper is to study missing value imputation in clinical contexts, two real-world medical datasets were chosen: the Osteoporosis Dataset and the Cardiothoracic Surgery Dataset.

The Osteoporosis Dataset assembles publicly available data from the 2013-2014 cycle of the NHANES (National Health and Nutrition Examination Survey Data, 2022). The dataset consists of 37 variables and 1643 subjects classified

into three conditions: normal, osteopenia, and osteoporosis. The osteopenia and osteoporosis classes were combined into a single one, thereby transforming this case study into a binary classification problem. As for the collected data, Table 1 gives a brief characterisation of each feature group within the Osteoporosis Dataset, including the number and type of variables, and the average missing rate.

Table 1
Brief characterisation of the Osteoporosis Dataset.

Feature Group	Attributes	Average missing rate (%)
Demographics	1 nominal, 2 ordinal, 1 numeric	$(1.1 \pm 1.9)\%$
Nutrition	6 numeric	$(7.6 \pm 0.2)\%$
Blood pressure	2 numeric	$(3.3 \pm 0.3)\%$
Anthropometrics	2 numeric	$(0.6 \pm 0.1)\%$
Physical fitness	1 numeric	10.0%
Blood lipids	4 numeric	$(27.8 \pm 25.4)\%$
Hormones	3 numeric	$(8.9 \pm 4.5)\%$
Biochemistry	2 numeric	$(3.1 \pm 0.7)\%$
Physical activity	11 numeric	$(10.0 \pm 0.1)\%$
Lifestyle	2 numeric	$(9.8 \pm 0.0)\%$

After categorical encoding, the final working dataset was left with 43 variables, 36 numerical and 7 binary. The dataset has a proportion of 73.2% incomplete samples, i.e. subjects with at least one missing value. Furthermore, 38.8% of the individuals were classified as normal (healthy), and the remaining 61.2% were diagnosed with either osteoporosis or osteopenia. The average age of all participants was 58 ± 12 years, with those classified as normal having a mean of 58 ± 10 years and the remaining, considered not healthy, an average age of 62 ± 12 years.

As for the Cardiothoracic Surgery Dataset, it contains clinical and demographic information retrieved by the Cardiothoracic Surgery Service of Hospital de Santa Marta in Portugal from 2011 to 2019. For each subject, the collection started during the pre-surgery period and extended up to one year after the surgical procedure. The dataset contains records from 8122 patients and was used to predict the occurrence of complications within three months after hospital discharge, in a binary classification problem. Data was collected on 106 medically relevant variables, which can be grouped into the categories shown in Table 2.

Table 2
Brief characterisation of the Cardiothoracic Surgery Dataset.

Feature Group	Attributes	Average missing rate (%)
Hospitalisation	4 dates	$(23.7 \pm 47.2)\%$
Cardiac history	4 ordinal, 1 binary	$(16.1 \pm 35.3)\%$
Previous interventions	2 dates, 1 ordinal, 1 nominal	$(44.9 \pm 51.6)\%$
Pre-operative risk factors	4 numeric, 3 ordinal, 4 nominal, 4 binary	$(0.9 \pm 0.2)\%$
Pre-operative haemodynamics and catheterisation	1 date, 5 numeric, 2 ordinal, 1 nominal, 1 binary	$(49.3 \pm 42.8)\%$
Pre-operative status and support	4 binary	$(0.6 \pm 0.2)\%$
Operation	1 text, 2 ordinal, 4 nominal	$(27.2 \pm 46.1)\%$
Coronary surgery	2 numeric, 2 nominal	$(61.4 \pm 0.7)\%$
Valve surgery	1 code, 8 numeric, 3 ordinal, 10 nominal, 7 binary	$(34.8 \pm 30.2)\%$
Cardiac Surgery Morbidity Scale	1 numeric, 8 nominal	$(7.6 \pm 22.1)\%$
Discharge details	2 nominal, 1 binary	$(32.4 \pm 55.6)\%$
Patient demographics/ autocalculations	1 code, 10 numeric, 1 nominal, 1 binary	$(7.8 \pm 26.4)\%$

An initial pre-processing was performed, and the working dataset was left with 5625 subjects, 92.7% of which are negative cases. Those negative cases have an average age of 65 ± 13 years, whereas the positive cases have an average

age of 68 ± 12 years. After applying categorical encoding, the resulting number of features is 119, of which 31 are numeric, and 88 are binary. Finally, 94.1% of the samples are incomplete.

In order to perform a time-based analysis, patients whose information was collected in 2019 were assembled in a separate test subset. This group contains 667 subjects, from which 92.7% are negative cases. Therefore, the distribution of class labels is maintained in this test set.

5.2. Experimental Setup

Prior to imputation, a grouped stratified 5-fold strategy was applied to each dataset. Note that the UCI Machine Learning Repository datasets were divided after injecting synthetic missing values.

The performance of the proposed imputation techniques was evaluated through a comparative study. Overall, seven other imputation methods were selected to be part of this study: Mean / Mode, Regression, KNN, CMIM (Sefidian and Daneshpour, 2020), CoHiKNN (Liu et al., 2019), NMVI (Bhagat and Singh, 2022), and MICE (Van Buuren and Groothuis-Oudshoorn, 2011). CMIM are a compilation of ten distinct imputation techniques; for this work only the fifth one was implemented and tested. Additionally, the results produced by listwise deletion were also included in the performed analysis. This selection sought to encompass both standard and modern methods with diverse baseline strategies. These methods served as benchmarks, enabling a more informative evaluation of the proposed techniques.

Two types of performance evaluation were carried out: imputation precision and classification evaluation. The first type assessed the quality of the imputation procedure by comparing the imputed values with the original ones, i.e. ground truth, resorting to the mean absolute error (MAE). This evaluation could only be carried out on the incomplete datasets generated from the three UCI Machine Learning Repository datasets since it is impossible to trace back the real values of the missing elements in the remaining datasets. The classification evaluation studied the impact of each imputation procedure on an ML model's performance, namely on RF, SVM, and NB classifiers. A 5-fold cross-validation strategy was adopted, and the average AUROC was used to compare the various imputation techniques.

The performed comparative study aimed to be as rigorous and fair as possible. To this end, hyperparameter tuning constituted an essential step to guarantee that each imputation technique was being evaluated under the best possible conditions, minimising the likelihood of any factors external to the imputation procedure corrupting the results. A grid search technique was applied to each stratified fold of the original UCI Machine Learning Repository databases, instead of the imputed datasets. The procedure followed for the real-world datasets was slightly different. As there is no baseline dataset (ground truth), it was considered necessary to compute a grid search with 5-fold cross-validation for each stratified fold of each imputed dataset to ensure a fair comparison between the imputation techniques. In order to soften the computational cost, the number of tested values within some hyperparameters of the RF classifier was lowered, as shown in Table 3. For each classifier, the optimal hyperparameters are the ones from the model with the highest average AUROC. The AUROC was therefore selected to assess each classifier's performance.

6. Results and Discussion

The imputation quality of the proposed methods will be first discussed. The average value of the three distinct MAEs corresponding to the three chosen missing rates was computed for each imputation method when applied to every synthetic dataset, as displayed in Table 4. Techniques unable to perform imputation for higher missing rates are marked, as the calculated errors do not fully represent their efficacy. Since the MAE frequently grows with the missing rate, because the amount of useful information decreases, the quality of the imputation performed by those techniques is likely worse than what the displayed MAEs indicate. For this reason, although the marked techniques sometimes have the lowest MAEs, they cannot be considered the best in terms of imputation quality.

The proposed correlation-based imputation methods yield consistently good results in all three missingness mechanisms. Notably, compared to its competitors, CWKNNI is the most precise technique in the MCAR mechanism. Furthermore, the ranking concerning the quality of the MCAR imputation procedure remains nearly unchanged independently of the dataset, i.e. techniques that are superior (inferior) in the Wine Data Set are also superior (inferior) in the SPECT Heart Data Set and the Statlog (Heart) Data Set. Figure 1 provides a visual representation of this observation. Hence, for MCAR data of both numeric and binary types, it is inferred that CWKNNI will produce estimates that are closer to the real values if they had been observed.

As for the MAR mechanism, the proposed methods exhibit an imputation quality close to their competitors, although not clearly better. The injection of MAR values was based on the pairing of highly correlated features, where one of the features determines the missing elements in the other. Hence it was expected that the proposed

Table 3
Tested hyperparameter values for every classifier.

Classifier	Hyperparameter Values
UCI Machine Learning Repository datasets	
RF	**max_depth** $\in \{3, 5, 7, 9\}$, **n_estimators** $\in \{5, 10, 25, 50, 100, 200\}$, **criterion** $\in$ {'entropy', 'gini'}, **min_samples_split** $\in \{10, 20, 50\}$, **min_samples_leaf** $\in \{5, 10, 25\}$
SVM	**random_state** $= 42$, **class_weight** $=$ 'balanced', **C** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$, **gamma** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$
NB	**var_smoothing** $\in \{1.0 \times 10^{-4},\ 2.8 \times 10^{-5},\ 7.7 \times 10^{-6},\ 2.2 \times 10^{-6},\ 6.0 \times 10^{-7},\ 1.7 \times 10^{-7},\ 4.6 \times 10^{-8},\ 1.3 \times 10^{-8},\ 3.6 \times 10^{-9},\ 1.0 \times 10^{-9}\}$
Real-world medical datasets	
RF	**max_depth** $\in \{3, 6, 9\}$, **n_estimators** $\in \{10, 25, 100\}$, **criterion** $\in$ {'entropy', 'gini'}, **min_samples_split** $\in \{20, 50\}$, **min_samples_leaf** $\in \{10, 25\}$
SVM	**random_state** $= 42$, **class_weight** $=$ 'balanced', **C** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$, **gamma** $\in \{0.0001, 0.001, 0.01, 0.1, 1, 10, 100\}$
NB	**var_smoothing** $\in \{1.0 \times 10^{-4},\ 2.8 \times 10^{-5},\ 7.7 \times 10^{-6},\ 2.2 \times 10^{-6},\ 6.0 \times 10^{-7},\ 1.7 \times 10^{-7},\ 4.6 \times 10^{-8},\ 1.3 \times 10^{-8},\ 3.6 \times 10^{-9},\ 1.0 \times 10^{-9}\}$

methods would yield the best results out of all techniques since they account for the correlation between values in the imputation process. Although their imputation quality was among the best, this superiority was not observed. The impact of missingness on the correlation between values, which increases as the missing rate grows, may have harmed the precision of the imputation procedure.

The MNAR mechanism generally presents the largest MAEs in all imputation methods, which was expected given that most existing techniques, both standard and state-of-the-art, are MAR-based approaches, and thus provide better results under the MCAR and MAR mechanisms. As for the proposed methods, their MAEs are also higher on the MNAR mechanism and the values are similar to those yielded by the remaining imputation techniques.

Regarding leveraging correlation for the prediction of missing values, the quality of the imputation performed by the proposed techniques CWKNNI and KNNSCI can be considered overall superior to that of competing correlation-based methods, i.e. CMIM and CoHiKNN. Furthermore, unlike these methods, the developed techniques could perform imputations for every missing rate, which again shows their superiority.

Table 5 concerns the classification performance evaluation, in which three different ML classifiers (RF, SVM, and NB) were trained upon the imputed datasets, and their performance was compared in terms of AUROC. Recall that this evaluation also includes results concerning the imputation performed on the two real-world medical datasets. As for the synthetic datasets, the AUROCs regarding all missing rates of every mechanism were averaged. As before, techniques unable to perform imputation on certain datasets are marked.

For each classifier, the proposed correlation-based imputation techniques are comparable to their competitors and can even be considered superior in some cases. Particularly, RFs that were trained upon datasets imputed through the proposed KNNSCI method exhibit an overall better performance, presenting the best AUROCs in 4 out of 6 evaluations, including the two real-world datasets, as shown in Figure 2. Although this result is not statistically significant, combining the KNNSCI method with an RF classifier obtained the best classification performances in

Table 4
Average MAE for all missingness mechanisms. The highlighted values are the lower MAEs in each assessment.

Imputation Method	Wine	SPECT Heart	Statlog (Heart)
MCAR mechanism			
Mean / Mode	0.18 ± 0.00	0.37 ± 0.01	0.25 ± 0.01
KNN	0.14 ± 0.01	0.26 ± 0.04	0.17 ± 0.05
NMVI	0.18 ± 0.03	0.42 ± 0.02	0.29 ± 0.00
CMIM	$0.15^{(a)}$	$0.29^{(a)}$	$0.18^{(a)}$
CoHiKNN	$0.15^{(a)}$	$0.29^{(a)}$	$0.20^{(a)}$
Regression	$0.15^{(a)}$	$0.29^{(a)}$	$0.19^{(a)}$
MICE	0.17 ± 0.01	0.34 ± 0.05	0.23 ± 0.04
CWKNNI*	**0.13 ± 0.01**	**0.25 ± 0.03**	**0.16 ± 0.03**
KNNSCI*	**0.13 ± 0.01**	0.26 ± 0.04	**0.17 ± 0.03**
CWRI*	0.15 ± 0.00	0.29 ± 0.01	0.18 ± 0.02
MAR mechanism			
Mean / Mode	0.23 ± 0.01	0.46 ± 0.05	0.35 ± 0.08
KNN	**0.13 ± 0.03**	**0.19 ± 0.09**	0.21 ± 0.11
NMVI	0.14 ± 0.02	0.43 ± 0.06	0.32 ± 0.12
CMIM	0.19 ± 0.04	$0.23 \pm 0.03^{(b)}$	0.21 ± 0.08
CoHiKNN	0.19 ± 0.02	$0.16 \pm 0.01^{(b)}$	0.28 ± 0.10
Regression	0.19 ± 0.05	$0.13 \pm 0.01^{(b)}$	19.02 ± 20.38
MICE	0.15 ± 0.03	0.29 ± 0.05	**0.18 ± 0.03**
CWKNNI*	**0.13 ± 0.03**	0.22 ± 0.10	0.24 ± 0.11
KNNSCI*	**0.13 ± 0.03**	0.20 ± 0.09	0.23 ± 0.11
CWRI*	0.18 ± 0.02	0.25 ± 0.13	0.27 ± 0.07
MNAR mechanism			
Mean / Mode	0.35 ± 0.00	0.62 ± 0.11	0.41 ± 0.13
KNN	0.24 ± 0.05	0.38 ± 0.17	0.30 ± 0.11
NMVI	**0.16 ± 0.04**	0.49 ± 0.08	0.33 ± 0.05
CMIM	0.27 ± 0.06	0.41 ± 0.18	$0.18 \pm 0.02^{(b)}$
CoHiKNN	0.29 ± 0.02	0.41 ± 0.27	$0.30 \pm 0.04^{(b)}$
Regression	0.27 ± 0.07	$0.24 \pm 0.04^{(b)}$	$0.20 \pm 0.01^{(b)}$
MICE	0.24 ± 0.05	**0.35 ± 0.08**	**0.27 ± 0.03**
CWKNNI*	0.24 ± 0.05	0.39 ± 0.20	0.31 ± 0.12
KNNSCI*	0.24 ± 0.05	0.39 ± 0.16	0.30 ± 0.11
CWRI*	0.31 ± 0.01	0.45 ± 0.22	0.36 ± 0.07

[(a)] No average was computed and only the MAE for a missing rate of 10% is shown because the technique was unable to perform imputation for the missing rates of 30% and 50%.
[(b)] The average was only computed with the MAEs for the missing rates of 10% and 30% because the technique was unable to perform imputation for a missing rate of 50%.
* Algorithms developed in this work.

the two real-world medical datasets, which is a desired achievement when developing reliable and robust ML-based systems to deploy in clinical contexts.

Moreover, correlation-based imputation yields the best AUROCs in 10 out of 18 evaluations, with the proposed KNNSCI and CWRI techniques being the main contributors to these results. These imputation techniques not only produced higher AUROCs but were also capable of performing imputation under all tested circumstances.

Initially, we believed that a more accurate imputation entailed a better classification performance. In order to investigate this hypothesis, the ML classifiers were also trained upon the original (and complete) UCI Machine Learning Repository datasets. Since these datasets represent an optimal imputation quality, i.e. a null MAE, comparing the obtained results with those of the classifiers trained upon imputed datasets allows the hypothesis to be tested.

Table 5
Average AUROC(%) obtained for each classifier. The highlighted values are the higher scores in each assessment.

	Wine	**SPECT Heart**	**Statlog (Heart)**	**Osteoporosis**	**Cardiothoracic**	**Cardiothoracic (2019)**
RF classifier						
Original Dataset	100.00±0.00	81.04±10.47	91.47±3.90	N/A	N/A	N/A
Listwise Deletion	79.70±24.26[a]	70.04±11.94[a]	91.38±1.49[b]	79.96±5.59	77.55±14.35[d]	56.13±7.31
Mean / Mode	98.71±0.94	83.77±8.26	89.18±2.22	83.43±3.31	70.18±4.66	61.24±0.97
KNN	98.45±1.07	81.12±7.12	89.05±2.62	83.45±3.20	71.08±4.94	61.17±1.81
NMVI	98.27±2.05	83.86±6.03	88.06±2.59	78.04±2.59	65.61±5.04	57.07±0.94
CMIM	99.24±0.39[a]	86.50±5.50[a]	89.23±1.68[b]	83.36±3.36	[c]	[c]
CoHiKNN	98.85±0.51[a]	88.93±7.27[a]	89.30±1.44[b]	83.57±3.12	[c]	[c]
Regression	98.90±0.50[a]	81.91±5.03[a]	88.11±3.22[b]	83.21±3.20	70.36±4.81	60.16±0.97
MICE	**98.99±0.69**	83.02±4.83	89.12±1.96	83.37±2.92	70.63±4.51	59.60±1.41
CWKNNI*	98.56±0.91	82.25±5.19	89.33±1.71	83.64±2.85	[c]	[c]
KNNSCI*	98.57±0.98	**84.63±5.58**	89.09±2.07	**83.73±3.13**	**71.32±4.78**	**61.86±1.00**
CWRI*	98.72±0.56	83.12±6.26	**89.50±2.26**	83.07±3.66	71.10±1.92	60.23±1.39
SVM classifier						
Original Dataset	100.00±0.00	83.39±9.43	91.17±3.17	N/A	N/A	N/A
Listwise Deletion	91.97±15.99[a]	58.39±24.53[a]	89.55±2.9[b]	82.58±4.33	46.37±18.97	49.00±8.66
Mean / Mode	99.27±0.60	75.71±24.30	88.37±1.95	82.78±2.40	68.58±2.89	55.95±2.55
KNN	99.27±0.62	79.44±13.16	88.27±2.32	82.75±2.35	68.88±2.70	55.97±2.44
NMVI	98.78±1.72	**84.72±4.36**	88.31±2.09	80.39±3.19	68.54±2.92	55.61±2.56
CMIM	99.54±0.34[a]	85.42±3.88[a]	88.81±1.35[b]	82.57±2.52	[c]	[c]
CoHiKNN	99.50±0.36[a]	85.47±5.84[a]	89.26±1.09[b]	82.74±2.32	[c]	[c]
Regression	99.41±0.49[a]	83.85±3.68[a]	87.39±3.90[b]	82.29±2.25	63.72±4.18	**56.67±1.64**
MICE	99.35±0.54	80.77±14.89	88.49±2.00	82.62±2.59	**68.96±2.54**	55.56±2.89
CWKNNI*	99.35±0.50	80.91±9.92	88.54±1.46	82.91±2.70	[c]	[c]
KNNSCI*	99.23±0.59	83.35±6.00	88.35±2.45	**83.02±2.58**	68.91±2.68	56.14±2.63
CWRI*	**99.40±0.44**	84.32±5.76	**88.91±1.84**	82.78±2.41	68.58±1.85	55.93±3.24
NB classifier						
Original Dataset	99.82±0.25	74.86±4.86	86.50±4.33	N/A	N/A	N/A
Listwise Deletion	82.30±20.3[a]	68.97±9.65[a]	82.12±6.69[b]	77.28±5.35	50.85±14.71	49.58±3.16
Mean / Mode	97.79±1.90	74.63±5.95	85.50±1.55	77.54±2.82	66.58±4.57	58.00±0.88
KNN	98.21±1.13	75.92±4.04	85.93±1.99	77.80±2.93	67.05±4.27	57.86±0.78
NMVI	97.57±2.21	74.32±3.55	86.07±1.81	77.54±2.62	66.90±4.36	58.00±0.85
CMIM	98.07±1.48[a]	77.97±5.04[a]	85.32±1.42[b]	77.47±2.60	[c]	[c]
CoHiKNN	98.58±0.81[a]	78.01±5.85[a]	85.84±0.61[b]	77.58±2.93	[c]	[c]
Regression	98.00±1.50[a]	74.45±3.16[a]	85.69±2.37[b]	76.51±2.56	50.80±7.18	49.56±3.97
MICE	**98.45±1.04**	75.44±2.72	**87.38±1.66**	77.59±2.69	**67.45±4.07**	**58.04±1.15**
CWKNNI*	98.28±1.08	**77.36±5.70**	86.63±0.91	**77.90±2.89**	[c]	[c]
KNNSCI*	98.30±1.15	76.60±3.74	86.19±1.46	77.79±3.05	67.06±4.26	57.94±0.79
CWRI*	98.12±1.40	75.11±5.13	86.28±1.57	77.55±2.80	65.07±3.50	57.56±0.68

[a] The technique was unable to perform under the MCAR mechanism with missing rates of 30% and 50%.
[b] The technique was unable to perform under the MCAR mechanism with missing rates of 30% and 50%, and under the MNAR mechanism with a missing rate of 50%.
[c] The technique was unable to perform imputation due to high computational costs.
[d] Only 7.3% of the samples from the Cardiothoracic Surgery Dataset were used in this complete-case analysis, which is not at all representative.
* Algorithms developed in this work.

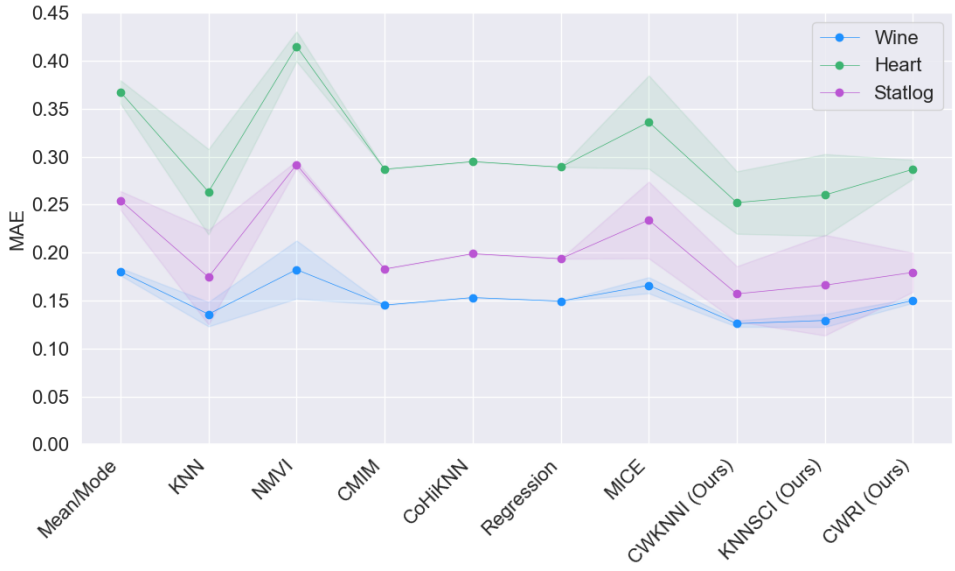


Figure 1: MAE for all synthetically generated datasets under the MCAR missingness mechanism.

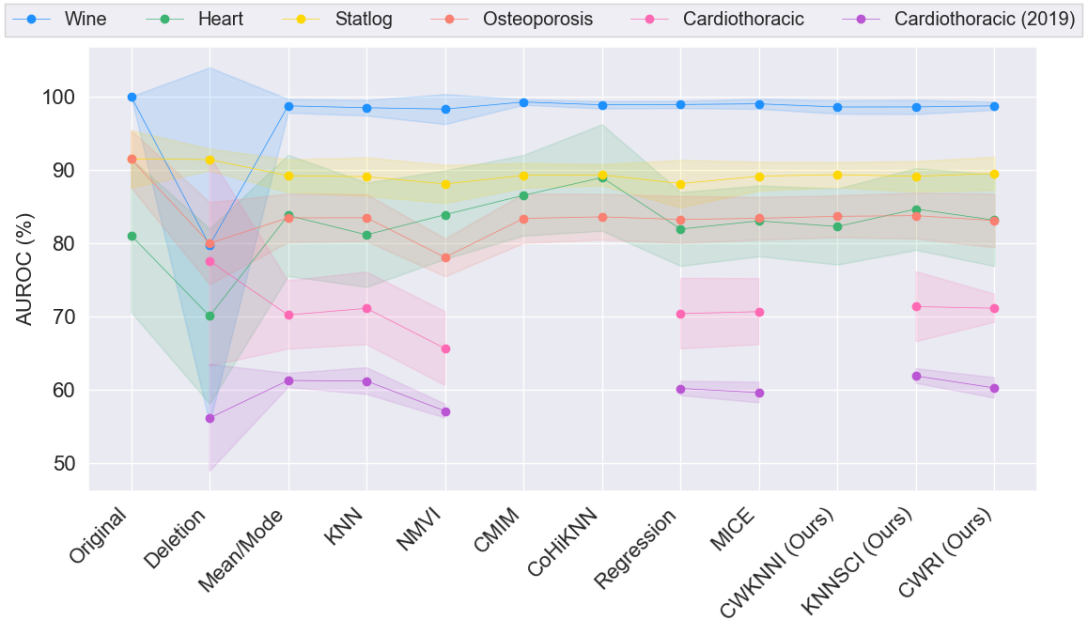


Figure 2: AUROC for every imputation method performed on the working datasets, obtained through an RF classifier.

Furthermore, recall that the classifiers' hyperparameters were chosen after a grid search was applied to the original dataset, and no hyperparameter tuning was performed on the models trained upon the imputed datasets. Hence, the classifiers trained upon the original datasets may have a slight advantage over the remaining, as their performance was optimised.

Even so, in the SPECT Heart Data Set, for example, the vast majority of the RFs trained upon imputed datasets outperform the RF trained upon the original dataset, thus demonstrating that there is not a clear relationship between imputation quality and the performance of an ML model. Figure 3 depicts this example. This finding raises the question of whether a more suitable imputation method for an ML-based clinical system is one that yields more precise estimates or one by which a better classification performance is achieved. At first glance, a clinical prediction model should have optimal performance, but it may not be acceptable to fully disregard the imputation quality of the chosen method. For instance, consider an imputation technique that produces biased parameter estimates, i.e. distorts the original statistical distribution of the data. In some cases, this permits a greater generalisation ability of the ML model trained upon the

imputed data. However, the knowledge that should have been learned from the data may have been corrupted by the imputation procedure, ultimately leading to a facilitated learning task and misleadingly better classification results.

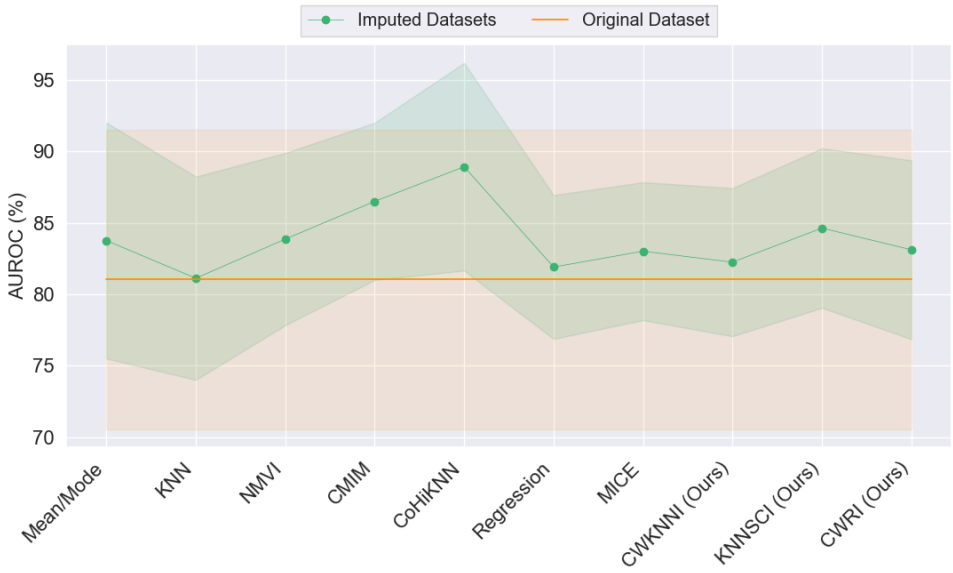


Figure 3: AUROC for every imputation method performed on the SPECT Heart Data Set, obtained through an RF classifier. These results are compared against the AUROC of the RF trained upon the original dataset (orange line).

Lastly, Table 5 shows that the most prominent differences in AUROC come from the scores produced by Listwise Deletion in comparison to the others. This reinforces that this approach should be used cautiously because simply discarding all incomplete instances may produce a dataset that is not representative of the original problem, particularly for higher missing rates.

7. Conclusions

Missing data are ubiquitous in biomedical sciences, posing a recurring predicament in delivering reliable AI-based clinical systems. There has been a growing interest in strategies that address this inevitable challenge, specifically missing value imputation methods.

This paper proposed three novel correlation-based imputation techniques which leverage not only the correlation between values but also the correlation between values and missingness patterns. Their performance was evaluated in a comparative study which included existing methods, both standard and state-of-the-art. This study assessed the imputation quality of the proposed methods under diverse missingness conditions and on distinct variable types. Furthermore, classification performance was assessed on multiple datasets, both synthetic and real-world. Hence, a comprehensive evaluation that is often lacking in literature was ensured. The proposed techniques were in compliance with their competitors, sometimes outperforming them. In fact, the best AUROCs for real-world medical datasets were obtained through an RF trained using data imputed with the proposed KNNSCI method.

One of the proposed imputation methods, CWKNNI, could not be applied to the most complex dataset due to its high computational cost. Even though computational cost should not be a determining factor for dismissing an imputation method, it is a concern that must be considered in future works. Moreover, note that this drawback is shared by some state-of-the-art techniques, although it was overcome by the proposed methods KNNSCI and CWRI.

We found that a more accurate imputation does not entail a better classification performance, which may imply that a trade-off between these two properties has to be made when choosing an imputation technique.

In summary, this work confirmed the auspicious role of correlation-based imputation in improving ML-based clinical systems' robustness to missing values while addressing important limitations of current imputation methods.

8. Acknowledgements

This work was done under the project "CardioFollow.AI: An intelligent system to improve patients' safety and remote surveillance in follow-up for cardiothoracic surgery", and supported by national funds through 'FCT – Portuguese Foundation for Science and Technology, I.P.', with reference DSAIPA/AI/0094/2020.

CRediT authorship contribution statement

Isabel Curioso: Conceptualization, Methodology, Software, Data Curation, Writing - Original Draft. **Ricardo Santos:** Conceptualization, Methodology, Validation, Resources, Data Curation, Writing - Review & Editing, Supervision. **Bruno Ribeiro:** Conceptualization, Methodology, Validation, Writing - Review & Editing. **André Carreiro:** Writing - Review & Editing. **Pedro Coelho:** Resources, Supervision. **José Fragata:** Resources, Supervision. **Hugo Gamboa:** Writing - Review & Editing, Supervision.

References

Akoglu, H., 2018. User's guide to correlation coefficients. Turkish journal of emergency medicine 18, 91–93. doi:10.1016/j.tjem.2018.08.001.

Ambinder, E.P., 2005. Electronic Health Records. Journal of oncology practice 1, 57. doi:10.1200/jop.2005.1.2.57.

Bhagat, H.V., Singh, M., 2022. NMVI: A data-splitting based imputation technique for distinct types of missing data. Chemometrics and Intelligent Laboratory Systems 223, 104518. doi:10.1016/j.chemolab.2022.104518.

Enders, C.K., 2022. Applied Missing Data Analysis. Guilford Publications.

Iranfar, A., Arza, A., Atienza, D., 2021. ReLearn: A Robust Machine Learning Framework in Presence of Missing Data for Multimodal Stress Detection from Physiological Signals. URL: https://arxiv.org/abs/2104.14278, doi:10.48550/ARXIV.2104.14278.

Kang, M., Tian, J., 2018. Machine Learning: Data Pre-processing. Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things , 111–130doi:10.1002/9781119515326.ch5.

Khan, H., Wang, X., Liu, H., 2022. Handling missing data through deep convolutional neural network. Information Sciences 595, 278–293. doi:10.1016/j.ins.2022.02.051.

Little, R.J., Rubin, D.B., 2019. Statistical Analysis with Missing Data. volume 793. John Wiley & Sons.

Liu, X., Lai, X., Zhang, L., 2019. A Hierarchical Missing Value Imputation Method by Correlation-Based K-Nearest Neighbors, in: Proceedings of SAI Intelligent Systems Conference, Springer. pp. 486–496. doi:10.1007/978-3-030-29516-5_38.

Mishra, P., Mani, K.D., Johri, P., Arya, D., 2021. FCMI: Feature Correlation based Missing Data Imputation. arXiv preprint arXiv:2107.00100 doi:10.48550/ARXIV.2107.00100.

National Health and Nutrition Examination Survey Data, 2022. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). URL: https://www.cdc.gov/nchs/nhanes/index.htm.

Rockel, T., 2022. missMethods: Methods for Missing Data. URL: https://CRAN.R-project.org/package=missMethods. r package version 0.3.0.

Rubin, D.B., 1976. Inference and Missing Data. Biometrika 63, 581–592. doi:10.2307/2335739.

Schober, P., Boer, C., Schwarte, L.A., 2018. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia 126, 1763–1768. doi:10.1213/ANE.0000000000002864.

Sefidian, A.M., Daneshpour, N., 2020. Estimating missing data using novel correlation maximization based methods. Applied Soft Computing 91, 106249. doi:10.1016/j.asoc.2020.106249.

Van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate imputation by chained equations in R. Journal of statistical software 45, 1–67. doi:10.18637/jss.v045.i03.

A. Appendix

Table A.1
Tested parameter values for every imputation method within the comparative study.

Imputation Method	Parameter Values
Mean / Mode	**missing_values** = np.nan
Regression	–
KNN	**missing_values** = np.nan, **n_neighbors** $\in \{5, 10, 15\}$, **weights** = 'distance', **metric** = 'nan_euclidean'
CMIM	**percentage** $\in \{0.1, 0.5, 0.9\}$, **threshold** $\in \{0.1, 0.5, 0.9\}$
CoHiKNN	**n_neighbors** $\in \{5, 10, 15\}$
NMVI	–
MICE	**m** = 3, **maxit** = 10, **seed** = 42, **defaultMethod** = c("pmm", "logreg", "polyreg", "polr")
CWKNNI	**n_neighbors** $\in \{5, 10, 15\}$, **percentage** $\in \{0.2, 0.4, 0.6, 0.8\}$, **initial_fill** $\in \{\text{False}, \text{True}\}$, **update** $\in \{\text{False}, \text{True}\}$
KNNSCI	**n_neighbors** $\in \{5, 10, 15\}$, **percentage** $\in \{0.2, 0.4, 0.6, 0.8\}$, **initial_fill** $\in \{\text{False}, \text{True}\}$, **update** $\in \{\text{False}, \text{True}\}$
CWRI	**percentage** $\in \{0.2, 0.4, 0.6, 0.8\}$