Received: 11 November 2023 | Revised: 18 April 2024 | Accepted: 25 April 2024 | Published online: 30 April 2024

RESEARCH ARTICLE

Artificial Intelligence and Applications
yyyy, Vol. XX(XX) 1–5
DOI: 10.47852/bonviewJAIA42022045
BON VIEW PUBLISHING

Cardiac Disease Diagnosis Using K-Nearest Neighbor Algorithm: A Study on Heart Failure Clinical Records Dataset

Vitória S. Souza[1], Danielli A. Lima[1,*]

[1] *Laboratory of Computational Intelligence, Robotics and Optimization (LICRO), Federal Institute of Education, Science and Technology of Triângulo Mineiro (IFTM) Campus Patrocínio MG, Brazil*

Abstract: This article introduces an approach to diagnose heart diseases utilizing the K-Nearest Neighbor algorithm and diverse correlation filters for selecting the most pertinent attributes. Results high- light that meticulous filter selection enhances survival predictions in patients with heart diseases. Employing K = 5 and correlation filter C_F = 0.1, key attributes for classification were identified as anemia, high blood pressure, serum creatinine, and sex. Omitting the 'time' attribute led to information loss but was crucial to prevent biases and generalize predictions across various clinical scenarios. Utilizing these classification parameters, we designed an Android mobile application called "Heart Info System", functioning as an artificial intelligence service. It employs the K-Nearest Neighbor algorithm with optimal parameters to evaluate the probability of survival in the progression of heart disease. The main activity of the application retrieves data from a Firebase database. While the study results show promise, the accuracy of the application may be influenced by inaccurate or incomplete input data. Nevertheless, this application has the potential to improve the early detection of heart diseases, paving the way for life-saving interventions.

Keywords: K-Nearest neighbors, classification algorithm, machine learning, expert system, heart failure, correlation filter, artificial intelligence as a service

***Corresponding author:** Danielli A. Lima, Laboratory of Computational Intelligence, Robotics and Optimization (LICRO), Federal Institute of Education, Science and Technology of Triângulo Mineiro (IFTM) Campus Patrocínio MG, Brazil. E-mail: danielli@iftm.edu.br

1 Introduction

The heart is one of the most important organs in the human body, supporting life in humans. Being such an important organ, it requires great care, as the development of a cardiovascular disease, if not properly treated, or if there is a disagreement in the exams or even in the person themselves, an event of death can occur [Iwano and Little, 2013; Ishaq et al., 2021; Santos and Bittencourt, 2008]. Heart diseases are one of the leading causes of mortality worldwide. They can be developed over time or acquired through heredity. Often, what leads to the development of heart diseases is poor nutrition. They include a range of diseases that affect the heart and blood vessels, such as heart failure, coronary artery disease, cardiac arrhythmias, among others [Ahmad et al., 2017; Kim et al., 2019]. Given their impact on public health, early identification of these diseases is crucial to prevent complications and save lives. These conditions are often diagnosed through invasive tests, which can be uncomfortable and costly for the patient, as well as consume significant resources from the healthcare system [Chicco and Jurman, 2020; Chicco et al., 2021].

By applying artificial intelligence (AI) in the field of medicine, together with medical analysis, we can provide significant improvements in the discovery, treatment, and analysis of various diseases [Ahmad et al., 2017]. In this work we used a public database (DB) called Heart failure clinical records Data Set[1]. The data was collected from 299 patients with heart failure and applying machine learning (ML) algorithms to predict the survival of patients with cardiovascular diseases [Chicco and Jurman, 2020; Chicco et al., 2021]. Investing in information systems (IS)

[1]UCI Machine Learning Repository - Heart failure clinical records Data Set https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records.

based on linked open data is an opportunity to promote better information management and universal access to knowledge [Siqueira et al., 2017]. We can identify valuable data for prognosis studies and heart failure treatment, as well as identify risk factors and predictors of mortality in patients with this condition. In particular, supervised learning, where the algorithm is trained from a set of labeled data, is one of the most common methods in identifying patterns in medical data [Dornelas and Lima, 2023; Raikwal and Saxena, 2012].

Among the supervised learning methods, K-Nearest Neighbors (KNN) is a simple and effective technique that is based on identifying the k nearest points to a given point to classify it [Zhang et al., 2018]. In this study, we used the KNN technique varying $|K| = 15$ times and evaluated its accuracy in identifying heart diseases. In addition, we used 10 correlation filters to select the most relevant parameters for classification, which allowed for faster processing and a reduction in the number of tests per patient.

Furthermore, this study describes the implementation of a mobile application type of expert system for Android devices, called the Heart Info System (HIS), which aims to help patients and doctors make decisions in diagnosing heart diseases. According to [Neto et al., 2017], expert systems are now ubiquitous and can be easily formed by non-programmers using interoperable IS. They are being used to improve sustainability, health and other aspects of human life, and new business models are emerging around them. In our approach, HIS application, allows the user to enter data from blood tests and other exams, and uses the KNN algorithm to identify heart diseases. This study evaluates the effectiveness of KNN in identifying heart diseases, using different correlation filters and evaluation metrics to determine the minimum number of attributes needed for accurate diagnosis. The proposed approach aims to contribute to the improvement of public health by providing an accurate and efficient method for the early identification of heart diseases with the minimum number of tests necessary. In summary, this study has the potential to improve the effectiveness and efficiency of the diagnosis of heart diseases and thus contribute to the promotion of health and quality of life of patients.

In this context, this paper presents an innovative approach to the diagnosis of heart diseases, using the KNN algorithm and different correlation filters to select the most relevant attributes [Xing and Bei, 2019; Raikwal and Saxena, 2012; Zhuang et al., 2020]. We hope that these results can contribute to improvements in public health by offering a more accurate and efficient method for diagnosing heart diseases, reducing costs and discomfort for the patient. The results are discussed and analyzed based on the parameters collected by the systematic variations arising from the classification algorithms. Finally, based on experiments conducted using different measures for the evaluation of the supervised learning algorithm, we present HIS, a mobile application considered Artificial Intelligence as a Service (AIaaS), which uses the KNN algorithm to rank the chance of survival in case of heart disease progression.

2 Fundamentals and Related Works

In this section, we will outline the theoretical framework of the article, providing essential definitions and insights into heart disease. The focus will extend to a detailed exploration of the supervised machine learning paradigm and the KNN algorithm, a fundamental and efficient classification algorithm employed in this study. The chosen KNN algorithm is renowned for its simplicity and effectiveness in supervised learning. Furthermore, we will introduce the dataset central to this work—the Heart Failure Clinical Records Data Set, a repository of patient information from the Institute of Cardiology and Allied Hospital Faisalabad, Pakistan, during April-December (2015) [Ahmad et al., 2017]. To contextualize our study, we will also review pertinent works by authors who have utilized the Heart Failure Clinical Records dataset, offering a comprehensive overview of the existing literature in this domain.

2.1 Definitions about heart disease

There are numerous cardiovascular diseases, including heart failure, myocardial infarction, and arrhythmias. Many of these conditions can be prevented or effectively treated through early and accurate diagnosis. These diseases, affecting the heart, may lead to fatalities if not appropriately addressed. They typically develop over time due to factors like poor diet, lack of exercise, or stress [Chicco and Jurman, 2020]. Examples of factors that exacerbate and pose a risk for heart disease include high blood pressure [Stamler et al., 1989; Vasan et al., 2001; Haider et al., 2003], commonly known as hypertension, characterized by elevated blood pressure levels in the arteries [Brouwers et al., 2021]. Another example is heart failure, a condition marked by the heart's inability to meet the body's demands, resulting in restricted blood flow and congestion in the veins and lungs [Santos and Bittencourt, 2008]. Acute myocardial infarction is defined as myocardial necrosis resulting from the obstruction of a coronary artery,among other conditions [Chicco and Jurman, 2020].

With the growing importance of machine learning in the discovery of heart diseases, it has become possible to use supervised learning algorithms to develop models capable of accurately predicting the presence of heart diseases. Supervised learning involves the use of labeled data to train models that can learn to make accurate predictions based on new unlabeled data. Among supervised learning algorithms, KNN (K-Nearest Neighbors)

is one of the most common, offering high accuracy and ease of implementation. To optimize the performance of KNN in detecting heart diseases, researchers have explored different correlation filters to select the most relevant attributes for classification. These filters are used to reduce the dimensionality of the data set and provide more relevant parameters for analysis.

The goal of this work is to evaluate the performance of the KNN algorithm by varying the value of K (from 3 to 31) and using different correlation filters $\{1, 0.85, 0.75, 0.5, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05\}$ to select the most relevant attributes. The ultimate goal is to develop an accurate and efficient model for diagnosing heart diseases with minimal invasive exams and processing time. To evaluate the performance of the model, we use different evaluation metrics, including accuracy, precision, recall, specificity, sensitivity, and f-measure. These metrics assess the model's accuracy and pinpoint areas for enhancement, aiming to optimize its efficiency.

2.2 K-Nearest neighbor algorithm

KNN is a machine learning classification algorithm that can be used to classify new data points based on the distance between them and known data points [Cambronero and Moreno, 2006; Deekshatulu et al., 2013; Enriko et al., 2018]. The algorithm works by finding the K closest known data points to the new data point and then assigning it the most frequent class among those K points [Phyu, 2009]. The choice of the value K can significantly affect the algorithm's performance, as too low a value may result in wrong classifications due to outliers, while too high a value may result in incorrect classifications due to the inclusion of many data points that are not relevant to the class of the new data point.

One of the main advantages of KNN is its simplicity and ease of implementation, as well as not requiring a distribution hypothesis for the data. The algorithm can also be adapted to handle regression problems, by finding the mean or median of the K closest neighbors to predict the numerical value of the new data point. However, the performance of KNN can be affected by the presence of noise in the data and the fact that it is not very efficient on large datasets, as it needs to calculate the distance between the new data point and all known data points for each prediction. Besides that, KNN can handle data with complex patterns and does not require the data to be linearly separable [Anggoro and Kurnia, 2020]. Because KNN makes predictions based on the nearest neighbors, it allows for localized decision-making and robustness to outliers, making it suitable for artificial intelligence as a service in a mobile device application. This enables the application to provide accurate and efficient predictions, even when operating in dynamic and potentially noisy environments. Unlike many other machine learning algorithms, KNN requires a minimal training phase, making it suitable for online learning scenarios and situations where the data distribution may change over time.

To improve the performance of KNN, there are some techniques that can be applied, such as normalizing the data to ensure that all features have the same scale, using more appropriate distance metrics for the dataset, and applying feature selection techniques to remove redundant or irrelevant features. In addition, there are also variations of KNN that can be used to improve performance, such as weighted KNN, which takes into account the distance of data points when making predictions, and kernel KNN, which uses a kernel function.

The choice of the K-Nearest Neighbors (KNN) algorithm was deliberate due to its simplicity, effectiveness, and suitability for our problem domain. KNN is a well-established classification algorithm known for its simplicity and intuitive nature [Zhuang et al., 2020], making it particularly suitable for our study on heart disease diagnosis. Some clustering methods rely on the concept of the Euclidean distance, a measure commonly used to assess the similarity between data points [Liang et al., 2022; Hu et al., 2023]. Additionally, KNN does not make any assumptions about the underlying data distribution, which is advantageous given the complexity and variability of medical data. Moreover, KNN is non-parametric, allowing it to adapt flexibly to the dataset without imposing strict assumptions on the data [Xing and Bei, 2019]. These characteristics make KNN a suitable choice for our study, where we aim to accurately classify heart disease cases based on patient attributes.

2.3 Database description

The dataset "Heart failure clinical Records" available in the UCI Machine Learning Repository contains clinical information about 299 patients with heart failure, collected at the Institute of Cardiology and Allied Hospital Faisalabad-Pakistan during April-December (2015) [Ahmad et al., 2017]. There are a total of 13 attributes that describe the clinical and demographic characteristics of these patients, which are explained in detail below:

1. Age: indicates age of the patient in years (integer number).
2. Anaemia: indicates the presence of anemia in the patient (0 = no, 1 = yes).
3. High blood pressure: indicates the presence of hypertension in the patient (0 = no, 1 = yes).
4. Creatinine phosphokinase (CPK): biochemical parameter, which is the level of the CPK enzyme in the patient's blood.

5. Diabetes: indicates if the patient has diabetes (0 = no, 1 = yes).

6. Ejection fraction: percentage of blood that is ejected from the heart during each beat, which is an important measure of heart function.
7. Platelets: number of platelets in the patient's blood.
8. Sex: gender of the patient (0 = female, 1 = male).
9. Serum creatinine: level of creatinine in the patient's blood.
10. Serum sodium: biochemical parameter about the level of sodium in the patient's blood.
11. Smoking: indicates if the patient is a smoker (0 = no, 1 = yes).
12. Time: follow-up period in days since the diagnosis of heart failure.
13. Death event: target attribute that indicates whether the patient survived or died during the follow-up period (0 = yes, 1 = no).

These attributes were selected based on previous medical knowledge about risk factors and prognosis of heart failure. Based on these attributes, it is possible to investigate which characteristics are most relevant for the diagnosis and prognosis of the disease, as well as to develop predictive models to assist in the treatment and prevention of heart failure.

2.4 Related works

Among the various cardiovascular diseases, we have Acute Myocardial Infarction (AMI), characterized by a clot that blocks blood flow to the heart, as well as hypertension and rheumatic heart disease. We will use a database consisting of 299 patients, 105 women and 194 men, aged between 40 and 95 years old, and possessing 13 attributes. As [Chicco and Jurman, 2020] cites in their study, both had left ventricular systolic dysfunction and previous heartfailure.

Each attribute may or may not be relevant for predicting patient survival or death, among the 13 attributes, one of them is the DEATH_EVENT class, which is the classification class, where it will be classified whether the patient survived the heart disease or not. The other attributes are sex, anemia, creatine phosphokinase, serum creatinine, ejection fraction (EF), age, diabetes, high blood pressure, platelets, smoking, time, and serum sodium [Ahmad et al., 2017]. As one of the branches of machine learning, supervised learning is based on a model that can learn from predefined results, using well-labeled data. Thus, it can train the algorithm to perform a specific task [Zhang et al., 2018]. In this work, eight different techniques will be applied for predicting better results.

In the study by [Ahmad et al., 2017], a study is elaborated on the population of Pakistan with heart failure, estimating the survival and mortality rate. Using the ROC curve, it was possible to detect that at a longer follow-up time, 81% was obtained in relation to the death event, while in a short time, it can only recognize 77%. In the work of [Chicco and Jurman, 2020], the authors address only two clinical parameters for the approach of survival in patients with heart failure, using the same database as [Ahmad et al., 2017], which are serum creatinine and ejection fraction, where the construction of machine learning models was based on.

The study conducted by [Ishaq et al., 2021] aimed to predict patient survival using various categorization models, such as Decision Tree (DT), Adaptive Boost Classifier (AdaBoost), Logistic Regression (LR), Random Gradient Classifier (SGD), Random Forest (RF), gradient augmentation classifier (GBM), an additive tree (ETC), a Naive Bayes Gaussian classifier (G-NB), and a support vector machine (SVM). To overcome the class asymmetry problem, the synthetic minority noise oxidation (SMOTE) technique was applied. Moreover, the RF was used to train the machine learning model with key features. In comparison to the full feature set, the experimental results showed that ETC outperformed the other models, achieving an accuracy value of 0.9262 with SMOTE in predicting survival in cardiac patients. However, our work differs from this study in that we applied filters to reduce the parameter characteristics space, unlike the usage of all attributes in [Ishaq et al., 2021].

In [Kumar et al., 2021], the authors proposed an IoT-enabled framework named Cardiac Diagnostics and Demographic Identification (CDF-DI) Resource Systems, secured by Public Key Infrastructure (PKI), for identifying various heart disease features related to heart failure (HF). They employed statistical and motor fixation techniques to analyze secondary cardiac data attachment. Patients with HF often experience kidney problems that result in elevated levels of Serum Creatinine (SC) and Serum Sodium (SS). The Random Forest (RF) algorithm was used to identify key features related to long-suffering survival status, which included follow-up months, CS, ejection fractionation (EF), creatine phosphokinase (CPK), and platelets, achieving a 96% accuracy. The same algorithm was used to recognize five key features related to category recognition with a 94% accuracy.

Additionally, the fifth vital characteristics, including CPK, SC, subsequent month, platelets, and EF, were found to be significant predictors for the long-suffering age selvedge with a clarity of 96%. The Kaplan Meier graph revealed that the elite period mortality in the very advanced age. The proposed resources have the potential to impact clinical practice and assist medical professionals in recognizing the likely survival status of cardiac long-suffering. The key factors for long-suffering survival are recommended to be stored mainly in the following month, SC, EF, CPK, and platelet score.

In our investigation, although we attained a lower accuracy rate with KNN, our deliberate focus was on attributes accessible through blood tests. This approach not only aligns with cost-effectiveness for patients but also

contributes to reducing financial burdens on public health entities. Notably, the inclusion of the hospitalization length parameter was avoided as it is influenced by the characteristics of the public health system in each country or region, making it an unbiased parameter. Furthermore, we conducted an analysis incorporating a correlation filter in our Heart Info System application streamlining the required diagnostic exams for patients.

3 Proposed Methodology

In this section, we present two main aspects: the first involves an analysis using K-Nearest Neighbor and the application of correlation filters to identify a reduced set of parameters capable of yielding accurate model predictions. We explain the metrics used to determine the optimal correlation filter and k values. Additionally, in a second phase, we introduce a mobile application for Android of the Artificial Intelligence as a Service (AIaaS) type, designed to predict heart failure with a reduced number of attributes.

3.1 Search type

First, regarding the methodological approach, the research is considered quantitative [Rodrigues et al., 2007], as the attributes are statistically evaluated through data collection for heart diseases that is available online and publicly. Additionally, the nature of the research is considered applied, since data classification algorithms are applied in a structured and specific database [Fleury and da Costa Werlang, 2016]. The procedures adopted here were through experiments, in which $X = 10^2$ simulations are performed with the purpose of making the best precision and accuracy for the data set studied. The research is explanatory [Raupp and Beuren, 2006], as we aim to connect the attributes identified through the correlation filter to understand the causes and effects of survival or death of people with heart disease.

3.2 Procedure for classification

In this section we present a procedure for analyzing a dataset using the K-Nearest Neighbor (KNN) algorithm. The methodology involves dividing the dataset into a training set (70%) and a test set (30%), and normalizing, standardizing, and transforming the training and test data. The methodology also includes initializing the list of values K and the list of correlation filter values to be tested. Furthermore, the evaluation metrics for each combination of K and correlation filter are initialized, such as accuracy, Cohen's kappa, precision, recall, specificity, sensitivity, and f-measure. For each combination of K and correlation filter C_F, the correlation filter is applied to the training and test data to select the most relevant attributes.

Then, the KNN algorithm is run with the current value of K and the standardized training data, and the desired evaluation metric is selected. The training and testing process is repeated 100 times, randomly splitting the dataset into training and test sets in each iteration (cross-validation), and the average of the evaluation metrics for the current combination of K and correlation filter is calculated. The combination of K and correlation filter that maximizes the desired evaluation metric is selected for the final test. Finally, the process is redone, excluding the time attribute from the dataset, as time is not considered an exam and can be extended to all patients.

1. Divide the dataset into a training set (70%) and a test set (30%).
2. Normalize, standardize, and transform the training and test data of the dataset.
3. Initialize the list of KNN values $K = \{3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31\}$ to be tested.
4. Initialize the list of correlation filter values to be tested considering $C_F = \{1.0, 0.85, 0.75, 0.5, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05\}$.
5. Initialize the evaluation metrics for each combination of K and correlation filter (e.g., accuracy, Cohen's kappa, precision, recall, specificity, sensitivity, f-measure).
6. For each combination of K and correlation filter C_F, do the following:
   (a) Apply the correlation filter C_F to the training and test data to select the most relevant attributes.
   (b) Run the KNN algorithm with the current value of K and the standardized training data and select the desired evaluation metric.
      i. Classify based on the K nearest neighbors, which are chosen based on their Euclidean distance to the new example.
      ii. Assign to the new example the most frequent class among the K neighbors, which represents the predicted survival rate for that patient.
   (c) Repeat the training and testing process 100 times, randomly splitting the dataset into training and test sets in each iteration (cross-validation).
   (d) Calculate the average of the evaluation metrics for the current combination of K and correlation filter.

7. Select the K and correlation filter (C_F) combination that maximizes the desired evaluation metric (e.g., accuracy) and use that combination for the final test.
8. Redo the process, excluding the time attribute from the dataset, as time is not considered an exam and can be extended to all patients.

3.3 Correlation filters

A correlation filter is a technique used to select the most relevant features in a dataset. This technique evaluates the relationship between variables in the dataset by calculating the correlation coefficient between them [Liu et al., 2021; Yuan et al., 2020; Lima et al., 2021]. The correlation coefficient measures the strength and direction of the linear relationship between two variables, ranging from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.

The correlation matrix is a table that shows the relationship between each pair of variables in a dataset, where each cell in the table contains the value of the correlation coefficient between the corresponding variables [Liu et al., 2021; Yuan et al., 2020]. This matrix can be used to identify which variables are most correlated with each other. A high value in the cell indicates a strong correlation between the corresponding variables and suggests that they may be redundant or representing similar information. On the other hand, a low value indicates a weak correlation or no relationship between the corresponding variables, which suggests that they may be unique and provide different information.

Correlation filters are useful for reducing the dimensionality of the dataset by removing redundant or irrelevant variables, which can improve the performance of machine learning models such as KNN in predicting the survival rate of patients with heart failure, for example. In this work, a set of filters with values $C_F = 1.00, 0.85, 0.75, 0.50, 0.30, 0.25, 0.20, 0.15, 0.10, 0.05$ and $C_F = 10$ were applied. The appropriate selection of correlation filters can help improve the accuracy of models by reducing the number of features used for predic-tion, i.e., which parameters were considered most relevant for survival analysis [Dornelas and Lima, 2023; Souza and Lima, 2023]. For the 8 classification algorithms, Table 1 represents each filter and presents the attributes included in the selection, showing which columns were considered by each filter.

3.4 Evaluation criteria

In this section we will present the results obtained by the model through the classification algorithms. Thus, it is fundamental to understand the measures to verify if the algorithm obtained a good result in question.

3.4.1 Accuracy

Accuracy is measured through four parameters [Liu et al., 2021; Yuan et al., 2020; Japkowicz, 2006]: (i) False positive (FP) is when the result is negative but classified as positive, (ii) False negative (FN) when the result is positive but classified as negative, (iii) True positive (TP) when it is actually true, i.e., the number of deaths, and (iv) True negative (TN) when it is actually negative, i.e., how many did not die. By counting all these terms and obtaining the confusion matrix, it is possible to calculate evaluation metrics of accuracy (A) for classification, according to Equation 1, resulting in a value between $[0, 1]$.

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

The error is the cases where the algorithm could not get it right. In this sense, the error is calculated considering the difference of the total minus the accuracy value and is given by Equation 2.

$$E = 1.0 - A \tag{2}$$

3.4.2 Cohen's Kappa

Cohen's Kappa is a statistical metric used to evaluate the agreement between two or more annotations or classifications [Pérez et al., 2020; Warrens, 2015]. This metric is useful when there are more than two possible labels and when the class distribution is not uniform. The Cohen's Kappa coefficient (κ) is calculated as the proportion of observed agreement minus expected agreement, divided by 1 minus expected agreement. It ranges from $(-1 \leq \kappa \leq 1)$, where (1) indicates perfect agreement, (0) indicates agreement at chance level, and (-1) indicates perfect disagreement.

The formula for calculating the Cohen's Kappa coefficient is given by Equation 3:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

Table 1. Attributes list based on inclusion and exclusion by filters.

Correlation Filters	Included	Excluded	Atributtes Remained
Column Filter = none attribute filtered			
1.00	13	0	age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, DEATH EVENT
0.85			
0.75			
0.50			
0.30	12	1	age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, time, DEATH_EVENT
0.25			
0.20	11	2	age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, DEATH EVENT
0.15	7	6	age, anaemia, ejection fraction, high blood pressure, platelets, sex, DEATH EVENT
0.10	5	8	creatinine phosphokinase, serum sodium, sex, time, DEATH EVENT
0,05	4	9	age, serum sodium, smoking , DEATH EVENT
Column Filter = attribute 'time' filtered			
1.00	12	0	age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, DEATH EVENT
0.85			
0.75			
0.50			
0.30	11	1	age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, DEATH EVENT
0.25			
0.20			
0.15	7	5	anaemia, ejection fraction, high blood pressure, platelets, sex, serum_creatinine, DEATH EVENT
0.10	5	7	anemia, high_blood_pressure, serum_creatinine, sex DEATH EVENT
0,05	4	8	age, serum sodium, smoking , DEATH EVENT

Where P_o is the observed proportion of agreement and P_e is the expected proportion of agreement. P_o is calculated as the number of matches divided by the total number of ratings. P_e is calculated as the sum of the product of the proportions of each label in the first and second annotation.

The interpretation of the Cohen's Kappa coefficient varies depending on the application [Pérez et al., 2020]. A general rule is that values below 0.2 indicate weak agreement, between 0.2 and 0.4 indicate fair agreement, between
0.4 and 0.6 indicate moderate agreement, between 0.6 and 0.8 indicates strong agreement and above 0.8 indicates almost perfect agreement. The Cohen's Kappa coefficient is widely used in areas such as medicine and psychology,where agreement between different evaluators or observers is essential. Therefore, we will use it in our analysis, asit is a medical basis.

3.4.3 Sensitivity and specificity

Sensitivity and specificity are two common metrics used in evaluating machine learning algorithms, particularly in binary classification problems [Ishaq et al., 2021]. Sensitivity measures the proportion of positive examples that were correctly identified by the model, while specificity measures the proportion of negative examples that were correctly identified by the model. The sensitivity (Se) is defined by the Equation 4:

$$Se = \frac{TP}{(TP + FN)} \tag{4}$$

Where true positives are positive examples that were correctly identified by the model and false negatives are positive examples that were incorrectly identified as negative by the model. On the other hand, the specificity (Sp) is defined in Equation 5:

$$Sp = \frac{TN}{(TN + FP)} \tag{5}$$

Where true negatives are negative examples that were correctly identified by the model and false positives are negative examples that were incorrectly identified as positive by the model.

When evaluating a binary classification algorithm, it is important to consider both sensitivity and specificity [de Oliveira et al., 2020]. Depending on the context, one or the other may be more important. For example, in a test for a serious illness, sensitivity is often more important as it is more critical not to miss any positive cases. In an airport security test, on the other hand, specificity may be more important as it is more critical to minimize false alarms.

3.4.4 Recall and precision

Recall (also known as true positive rate) and precision (also known as positive predictive value) are two common metrics used to assess the quality of a binary classification model [Ahmad et al., 2017; Alvarez, 2002].

The recall measures the proportion of true positives (TP) that were correctly identified by the model in relation to the total number of positive samples ($TP + FN$). In other words, the recall measures the model's ability to correctly identify all positive cases. The formula for the recall (Rec) is given by Equation 6:

$$Rec = \frac{TP}{(TP + FN)} \tag{6}$$

Precision measures the proportion of true positives (TP) that were correctly identified by the model in relation to the total number of samples classified as positive ($TP + FP$). In other words, precision measures the model's ability to correctly identify all positive cases against the total number of cases that were classified as positive. The formula for the precision (Prc) is given by Equation 7.

$$Prc = \frac{TP}{(TP + FP)} \tag{7}$$

Both metrics range from 0 to 1, with values closer to 1 indicating a more accurate and effective model.

3.4.5 F-measure

The F-measure (also known as the F1 Score) is a measure that combines accuracy and recall into a single metric, providing an overall assessment of the effectiveness of a binary rating model [de Brito et al., 2020]. The F-Measure is calculated from the harmonic mean of precision and recall. The Equation 8 represents the F1-measure:

$$F1 = \frac{2 \times (Prc \times Rec)}{(Prc + Rec)} \tag{8}$$

The F1 Score is a measure that varies from 0 to 1, and the higher the value, the better the performance of the model. A model with an F1 Score of 1.0 is perfect, while a model with an F1 Score of 0.0 is completely ineffective.

3.5 Flowchart for the proposed machine learning

As shown in Figure 1, data were initially collected from the UCI Machine Learning Repository[2] and placed in Excel .xlsx format and we used the KNIME Analytics Platform to read data from 299 patients. In the data preprocessing phase, some columns with string format data were transformed into numbers and numeric data were transformed into strings, which were then subject to a string replacer to perform the substitutions correctly. We then filtered out rows with inconsistent values for the age attribute and removed them. Next, we normalized each data column and then performed denormalization for some of the data visualization and statistics.

We left a node for column filtering (if necessary) and applied linear correlation, which calculates the correlation coefficient for each pair of selected columns, i.e., a measure of the correlation between the two variables. The correlation filter determines which columns are redundant (i.e., correlated) and filters them out. The output table will contain the reduced set of columns.

After performing all these steps, we applied the X-Partitioner, which represents a cross-validation loop. We reapplied the learning algorithm that was applied with a value of ($x = 10^2$) iterations. At the end of the loop, there must be an X-Aggregator to collect the results from each iteration. All nodes between these two nodes are executed as many times as iterations should be performed. We then chose the data mining algorithms for learning and prediction, in this case, KNN. Finally, the scorer is calculated, and a confusion matrix is calculated with the number of matches in each cell.

[2]Platform Heart failure clinical records Data Set, where contains the medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features https://doi.org/10.24432/C5Z89.

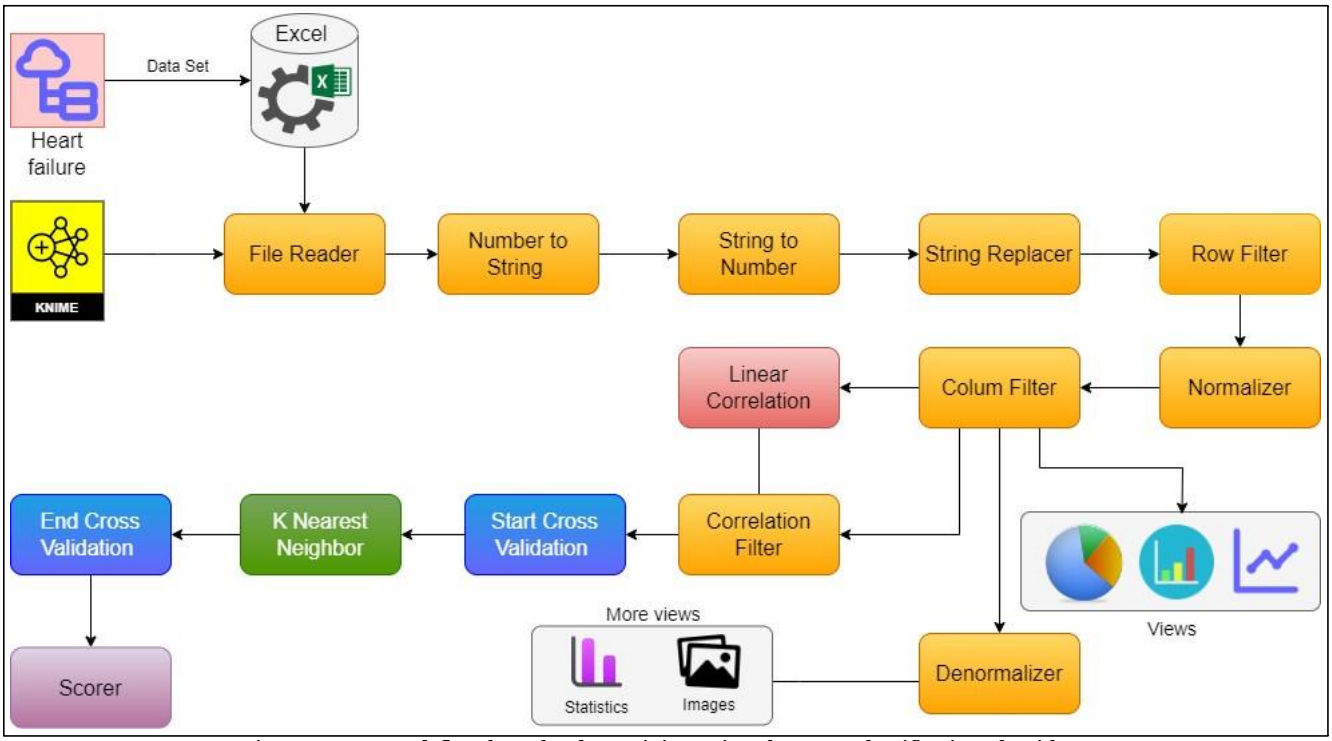


Figure 1. Proposed flowchart for data mining using the KNN classification algorithm.

3.6 Development in Knime analytics platform

To analyze the data we used the KNIME Analytics Platform, an open source platform and easy manipulation [Fillbrunn et al., 2017], [Dietz and Berthold, 2016]. In it, we created a workflow for classifying the database, as shown in Figure 2.

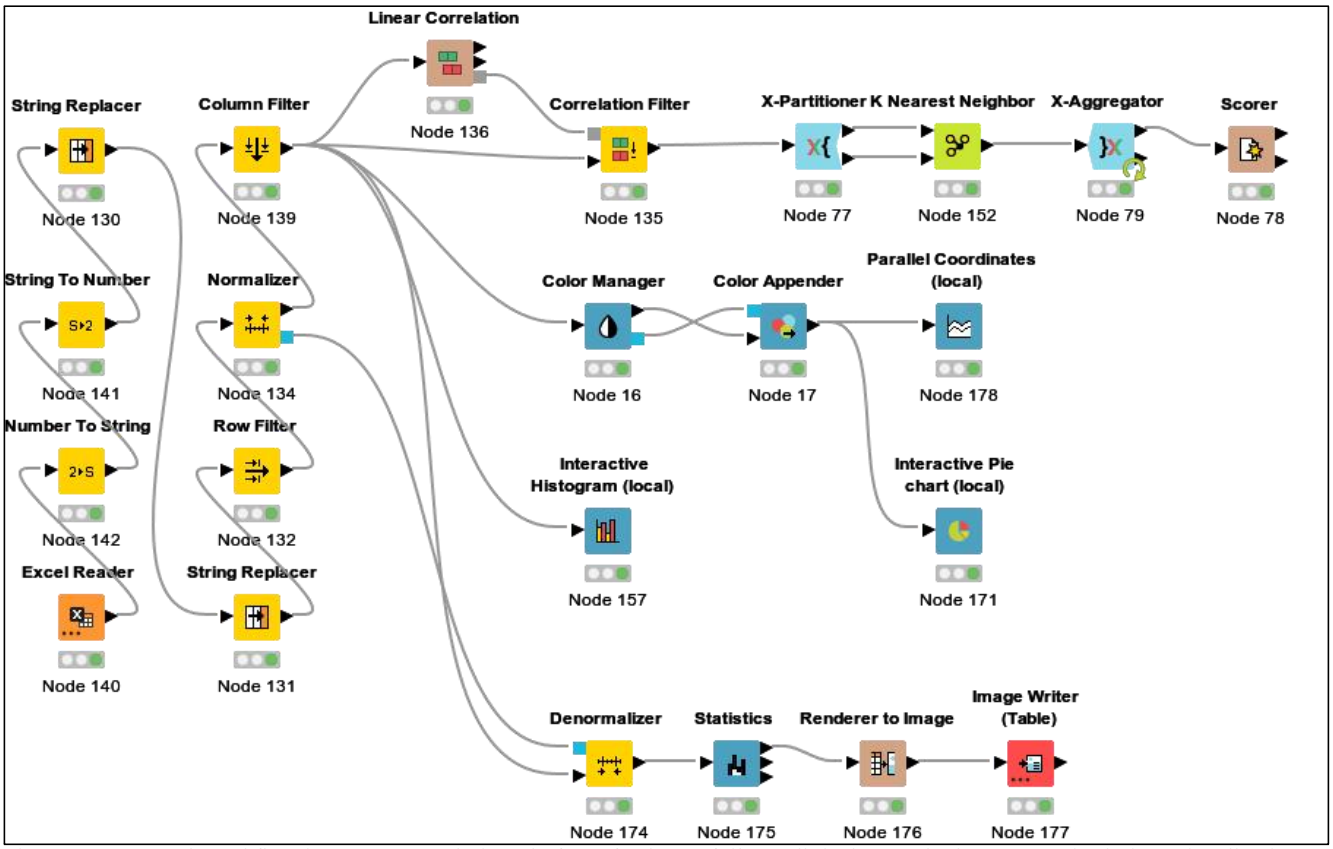


Figure 2. Proposed workflow in KNIME analytics platform for heart failure clinical records dataset manipulation, visualization,and data analytics.

In order to find a better result with the analysis of the data, we used filters in all algorithms. The filters are

used to demarcate which base columns will be more redundant for the evaluation, that is, which correlate, and consequently the relevant ones are filtered. As shown in the table (1), 10 correlation filters were used to make the dimensionality reduction of the parameters that need to be used in the data classification. The workflow was created in KNIME Analytics Platform and it follows the same flow as our proposal for data mining shown in Figure 1. Moreover, within KNIME, various nodes are dedicated to visual representation, including histograms illustrating attributes, providing a comprehensive visual analysis of the dataset.

3.7 Heart info system proposed

The expert system for mobile devices was implemented as an AIaaS for Android activity, that is, a screen that the user can interact with. This activity is called "Analyze" and is displayed on the mobile device screen. The code uses the Firebase library to access and manipulate data stored in a remote database. In particular, the code references two "keys" (myRefOrig and myRefNova) to retrieve data from the original base and the new base. The system is designed to display the result of data analysis on the screen. This result is displayed in a TextView called edtResTexto. The code defines a series of buttons that the user can click to perform different actions. For example, the btnHome button takes the user back to the main application screen, while the btnExit button closes the application completely.

The code defines a processData() method, retrieving and analyzing data from the database, utilizing custom classes like OriginalData and NewData. The user-friendly expert system incorporates features like action bars, custom colors for an appealing user interface (UI), and dialog boxes for displaying messages. Android features, XML layouts, and string resources are employed for UI, while external libraries like Kotlin handle specific tasks. Efficient data handling involves using structures like ArrayLists and programming techniques, including data filtering and array sorting.

Figure 3 shows how the back-end of the specialist Heart Info System (HIS) works for detecting heart diseases. The system uses a database (DB) called HFD database DB, where data collected from 299 patients with heart failure are stored. Initially, KNIME software is used to perform laboratory experiments in order to find the best

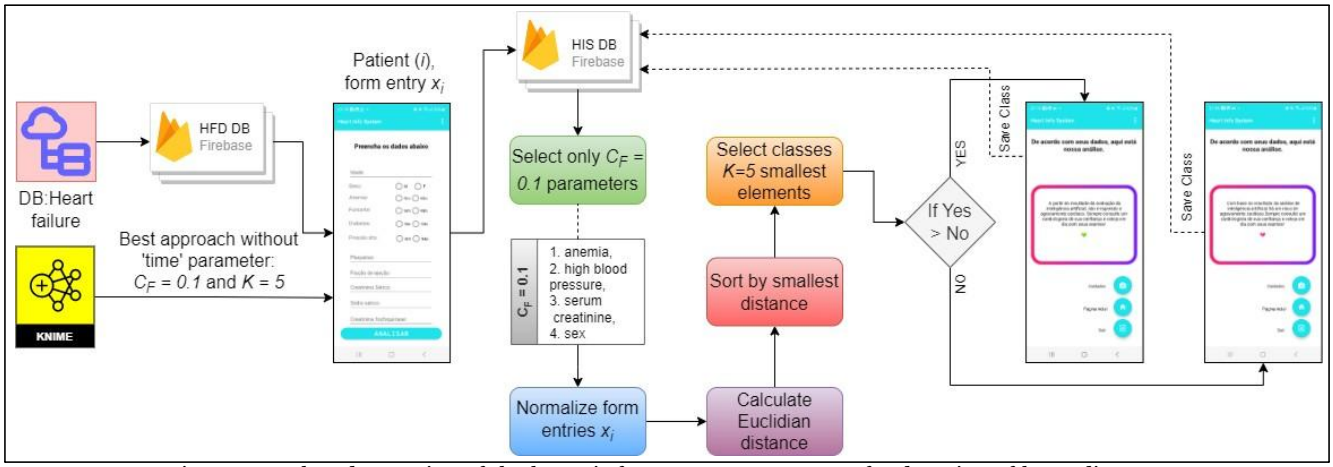


Figure 3. Back-end operation of the heart info system expert system for detection of heart diseases.

correlation filters (C_F) and the best value for the parameter K of the K-Nearest Neighbor (KNN) algorithm, which were determined as being 0.1 and 5, respectively. Next, a form is created that collects data from patient x_i and this data is normalized and stored in the HIS database.

Afterwards, the parameters that passed through the correlation filter (C_F) are selected and, using the Euclidean metric, the distance between the new data and the values already present in the HFD database is calculated. The closest K = 5 values are selected and the corresponding classes in HFD are accounted for. The most predominant class is attributed to the new patient x_i, the class is saved in HIS, remembering that this class has the accuracy returned by KNN.

The HIS is a tool that aids in diagnosing heart disease by analyzing patient data. It employs a machine learning algorithm, likely KNN based on the previous abstract, to assess the likelihood of a patient having heart disease. This information can be used by medical professionals to support their diagnosis and inform treatment decisions. An additional benefit of HIS is its ability to continuously improve its effectiveness. The system allows for updating data and parameters. This means that as new medical information and research becomes available, the HIS can be updated to reflect this knowledge. This continuous learning process helps the system maintain accuracy and provide the most up-to-date insights for better patient care.

Finally, it is important to remember that the KNN performance is contingent on dataset characteristics and patient profiles. Attribute selection in data analysis demands careful consideration of clinical relevance, data

availability, and problem-specific factors. Result validation across diverse datasets is essential for generalization. Exploring alternative machine learning techniques may enhance model performance.

4 Results

In this section, we will present a visualization of the data from the database under study. In this sense, we will present different types of graphs to understand the base that is being studied. Furthermore, based on the selected algorithms, we will present the classification results for each of the algorithms considering different correlation filters.

4.1 Data visualization

Firstly, each of the 12 attributes in the dataset was studied separately, as shown in Table 2 and Figure 4. According to the data in Table 2 and as can be seen in Figure 4(a), the age of the patients ranged from 40 to 95 years old, with a mean age of 64.43 years and median age of 60 years. The majority of patients (around 50%) are between 51 and 70 years old. For anemia (see Figure 4(b)), diabetes (see Figure 4(d)), high blood pressure (see Figure 4(f)), and smoking (see Figure 4(k)), the values range from 0 (absent) to 1 (present), with a mean of 0.43 for all patients. Around 57% of patients do not smoke or have anemia, diabetes, or high blood pressure. Creatinine phosphokinase is an attribute that measures the amount of creatinine phosphokinase in the patients' blood.

According to Table 2, this attribute has a minimum value of 23 and a maximum value of 7861, with a mean of 1438.29 and a median of 250 (see Figure 4(c)). There are extreme values for this attribute, with the highest value being almost 10 times greater than the third quartile. The ejection fraction attribute measures the percentage of blood ejected from the heart with each beat and had a minimum value of 14 and maximum of 80, with a mean of 40.86 and median of 38 (see Figure 4(e)). There is a large variation in the values for this attribute, with most patients (around 50%) having an ejection fraction between 30 and 45%.

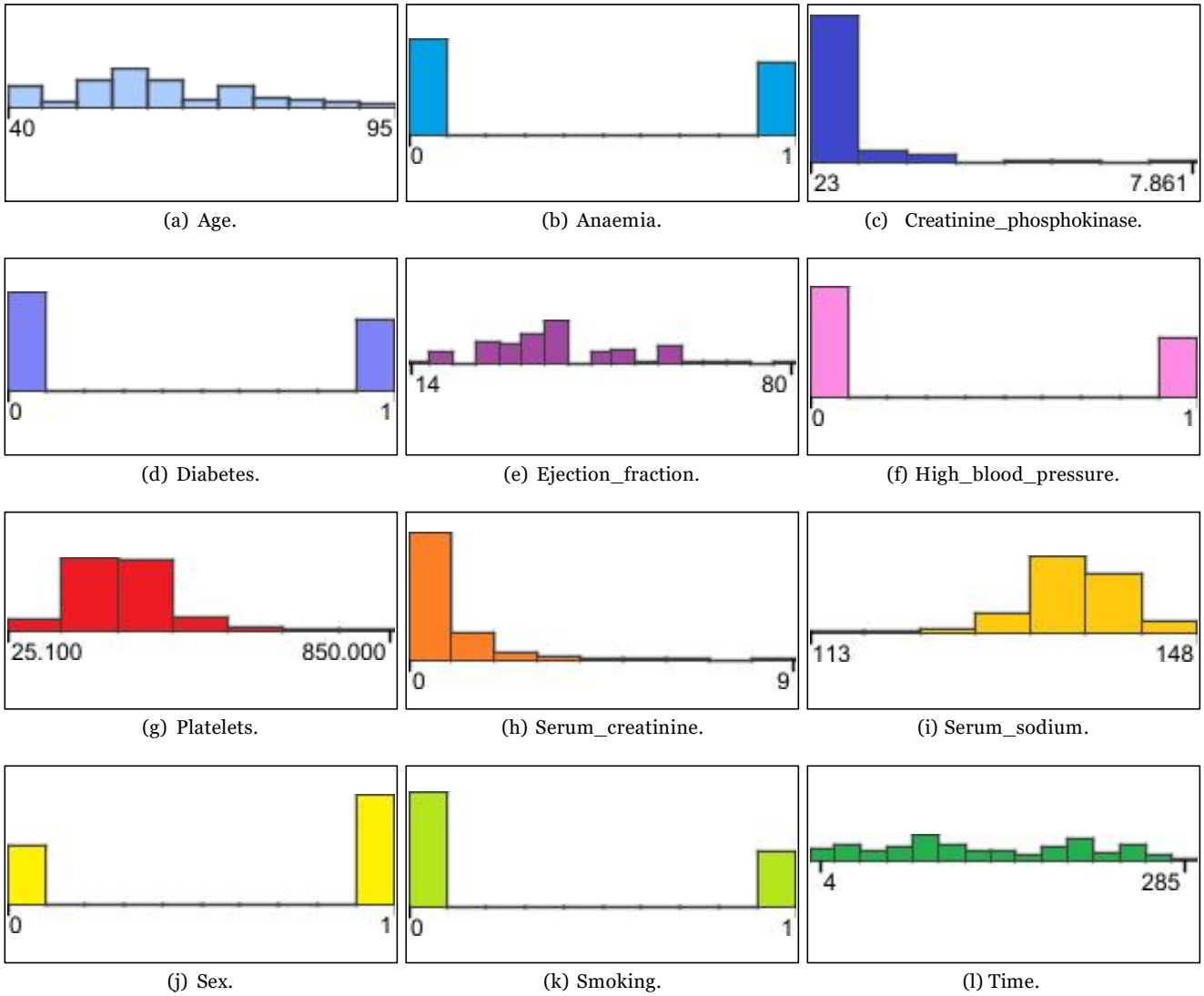
For the attribute platelets, which measures the amount of platelets in the patients' blood, the minimum value was 25100 and the maximum was 850000, with a mean of 309585.7 and a median of 262000. There are extreme values for this attribute, with the highest value being almost 3 times higher than the third quartile (see Figure 4(g)). As for the attribute serum creatinine, which measures the amount of creatinine in the patients' blood, the minimum value was 0.5 and the maximum was 9.4, with a mean of 2.27 and a median of 1.1. There are extreme values for this attribute, with the highest value being almost 3 times higher than the third quartile (see Figure 4(h)). The serum sodium attribute, which measures the amount of sodium in the patients' blood, had a minimum value of 113 and a maximum value of 148, with a mean of 135 and a median of 137 (see Figure 4(i)). The patients' sex was coded as 0 (female) and 1 (male), with a mean of 0.57, meaning that 57% of the patients are male (see Figure 4(j)).

The attribute time represents the time in days between follow-up and the start of the study. The median is 113 days, suggesting that half of the patients were followed up for less than 113 days and the other half for more than 113 days. The highest observation is 285 days, while the lowest is 4 days, indicating that the sample has a wide range of follow-up periods, as shown in Figure 4(l). In addition, the mean is 138.43 days, slightly higher than the median, suggesting that there may be some higher values pulling the mean upwards. The minimum value of 4 days suggests that patients were followed up for at least a few days, which is important for obtaining information on the progression of heart diseases.

In this context, the time attribute appears to be useful for understanding disease progression over time and can be used to predict the risk of cardiovascular events in patients with heart disease. The remaining attributes are presented in Table 2, such as the minimum and maximum value (considering outliers in the sample), lower and upper quartiles, median, the lowest and highest value for each attribute (disregarding outliers), and finally, the mean per attribute.

Table 2. Statistics for each of the attributes in the database.

Statistics	Age	Anaemia	Creatinine phosphokinase	Diabetes	Ejection fraction	High_blood pressure	Platelets	Serum creatinine	Serum sodium	Sex	Smoking	Time
Minimum	40	0	23	0	14	0	25100	0.5	113	0	0	4
Smallest	40	0	23	0	14	0	87000	0.5	125	0	0	4
Lower Quartile	51	0	118	0	30	0	213000	0.9	134	0	0	73
Median	60	0	250	0	38	0	262000	1.1	137	1	0	113
Upper Quartile	70	1	582	1	45	1	303000	1.4	140	1	1	205
Largest	95	1	1211	1	65	1	427000	2.1	148	1	1	285
Maximum	95	1	7861	1	80	1	850000	9.4	148	1	1	285
Mean	64.43	0.43	1438.29	0.43	40.86	0.43	309585.7	2.27	135	0.57	0.43	138.43



(a) Age. (b) Anaemia. (c) Creatinine_phosphokinase.

(d) Diabetes. (e) Ejection_fraction. (f) High_blood_pressure.

(g) Platelets. (h) Serum_creatinine. (i) Serum_sodium.

(j) Sex. (k) Smoking. (l) Time.

Figure 4. Frequency histograms for each database attributes.

4.2 Classification algorithm analysis

To evaluate the effectiveness of KNN, we used the Heart Failure database from the UCI Machine Learning Repository, which contains information about patients with heart failure.

We analyzed the KNN from two perspectives, initially considering accuracy: (i) with the parameter time (which informs the observation time of the patient in the hospital) and (ii) without the time parameter. Subsequently, we analyzed the database according to the other evaluation metrics for the KNN classification algorithm.

4.2.1 Analysing all parameters of dataset

Table 3 presents the results of simulations for the KNN, indicating that the performance of the algorithm varies significantly according to the value of K and the correlation filter used. Table 3 shows the results of the simulations for KNN, indicating that the algorithm's performance varies significantly according to the value of K and the correlation filter used. We observed that for values of K above 21, the accuracy starts to decrease, suggesting that selecting a relatively small ideal number of neighbors is important for the model's performance. The best performance found was 83.50% with a correlation filter of $C_F = 0.1$ and $K = 21$. The worst performance was 62.63% for $K = 5$ and a correlation filter of $C_F = 0.2$. In addition, the average accuracy for different values of K for each value of C_F also varied, with the best average accuracy (82.15%) obtained for the correlation filter of 0.1, while the worst performance (67.21%) was for the correlation filter $C_F = 0.2$. One possible explanation for these results is that the correlation filter may be removing important attributes for classifying patients, and that the choice of the ideal value of K may be influenced by the presence or absence of these attributes. In addition, the high variability in accuracy for different correlation filter values suggests that choosing a suitable filter may be challenging and should be carefully considered during the attribute selection and modeling process. Hence, other metrics will be

Table 3. Average accuracies and variations across different values of K in the KNN algorithm, along with correlation filters, considering the time attribute.

K	1.00	0.85	0.75	0.50	0.30	0.25	0.20	0.15	0.10	0.05
3	72.05%	72.05%	72.05%	72.05%	74.07%	74.07%	66.67%	69.70%	78.11%	68.35%
5	71.04%	71.04%	71.04%	71.04%	73.06%	73.06%	62.63%	67.68%	80.14%	67.68%
7	65.66%	65.66%	65.66%	65.66%	71.72%	71.72%	65.99%	67.34%	81.82%	69.02%
9	70.71%	70.71%	70.71%	70.71%	72.39%	72.39%	65.99%	67.68%	82.16%	69.02%
11	70.37%	70.37%	70.37%	70.37%	68.69%	68.69%	67.00%	66.67%	83.17%	68.69%
13	72.05%	72.05%	72.05%	72.05%	71.38%	71.38%	68.69%	67.34%	83.17%	68.01%
15	72.05%	72.05%	72.05%	72.05%	70.37%	70.37%	68.35%	66.67%	82.83%	69.70%
17	73.06%	73.06%	73.06%	73.06%	70.37%	70.37%	68.01%	67.00%	82.83%	70.71%
19	71.72%	71.72%	71.72%	71.72%	70.37%	70.37%	67.68%	68.01%	83.17%	70.37%
21	71.72%	71.72%	71.72%	71.72%	70.37%	70.37%	67.68%	64.98%	83.50%	70.03%
23	71.38%	71.38%	71.38%	71.38%	69.36%	69.36%	67.68%	67.68%	82.83%	70.37%
25	70.71%	70.71%	70.71%	70.71%	69.02%	69.02%	68.01%	69.02%	83.17%	70.03%
27	70.37%	70.37%	70.37%	70.37%	69.02%	69.02%	67.68%	68.01%	82.83%	69.36%
29	71.38%	71.38%	71.38%	71.38%	69.02%	69.02%	68.01%	68.35%	81.82%	69.70%
31	70.03%	70.03%	70.03%	70.03%	69.02%	69.02%	68.01%	68.35%	80.81%	70.03%
Mean	70.95%	70.95%	70.95%	70.95%	70.55%	70.55%	67.21%	67.63%	82.15%	69.41%
Std. Dev.	0.01676	0.01676	0.01676	0.01676	0.01659	0.01659	0.01494	0.01101	0.01468	0.00917

presented to help refine the final result for knowledge extraction.

4.2.2 Analysing parameters without time

We understand that the time attribute in the UCI Machine Learning database may be a relevant factor for the diagnosis of heart diseases. However, it is important to note that this attribute represents the time that the patient was under observation in a specific hospital, in this case, the Institute of Cardiology and Allied hospital Faisalabad-Pakistan [Ahmad et al., 2017]. As each country may have different protocols and hospitalization times, the observation time may vary greatly, and this may affect the generalization of the results in different contexts. Therefore, in this study, the time attribute was not considered in the analysis to avoid possible biases and to generalize predictions for different clinical scenarios. It is important to emphasize that attribute selection can be a crucial step in building machine learning models, especially in datasets with many attributes, as is the case with the UCI Machine Learning database. Thus, a careful analysis of attributes can help improve the effectiveness and generalization of models.

We performed simulations with different values of K and different correlation filters to evaluate the accuracy of the KNN in predicting survival rates in patients with heart disease. The results obtained are presented in Table 4, which displays the means and variations for each correlation filter value. It was found that, in general, an increase in the value of $K > 19$ resulted in a decrease in accuracy. However, the selection of an appropriate correlation filter with a value of $C_F = 0.1$ was important in increasing the model's accuracy. It was also observed that the variation in accuracy for different values of K was smaller for higher correlation filters. These results indicate that careful selection of correlation filters can lead to better survival predictions in patients with heart disease using KNN.

Based on the simulations carried out, we can observe that the performance of the KNN algorithm was influenced by the values of K and the correlation filters used. The best accuracy obtained was 73.06% for $K = 5$ and a correlation filter of 0.1, while the worst approach resulted in an accuracy of 62.63% for $K = 5$ and correlation filters 0.3, 0.25, 0.2. In addition, the analysis by average variation of the correlation filter showed that the filters 0.3, 0.25, 0.2 had the worst performance, with an average accuracy of 67.21%, while the correlation filter 0.05 had the best performance, with an average accuracy of 69.41%. Furthermore, it is important to point out that deleting the time attribute resulted in a significant performance loss for the diagnostics. These results suggest that it is important to carefully choose the value of K and the correlation filter to be used to maximize the performance of the KNN algorithm in predicting the survival rate of patients with heart failure.

4.2.3 Analyzing KNN by other metrics

Based on the experimental results performed, as shown in Table 5, the performance of the KNN was evaluated considering different values of K, correlation filters and evaluation metrics. The results showed that the performance of the KNN varied significantly as a function of these parameters. Observing the results, it is noticed that the KNN performance was better when the time variable was considered, especially for larger values of K (17 'with filter 1.0' and 21). This may be due to the fact that the time (length of stay and patient follow-up) variable can be an important factor in determining heart disease.

Table 4. Average accuracies and variations across different values of K in the KNN algorithm, along with correlation filters, without considering the time attribute.

K	1.0	0.85	0.75	0.5	0.3	0.25	0.2	0.15	0.10	0.05
3	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	66.67%	68.35%	70.71%	68.35%
5	67.34%	67.34%	67.34%	67.34%	62.63%	62.63%	62.63%	71.04%	73.06%	67.68%
7	66.67%	66.67%	66.67%	66.67%	65.99%	65.99%	65.99%	71.04%	70.37%	69.02%
9	68.35%	68.35%	68.35%	68.35%	65.99%	65.99%	65.99%	71.72%	66.67%	69.02%
11	65.99%	65.99%	65.99%	65.99%	67.00%	67.00%	67.00%	69.70%	69.70%	68.69%
13	69.02%	69.02%	69.02%	69.02%	68.69%	68.69%	68.69%	69.36%	69.70%	68.01%
15	68.69%	68.69%	68.69%	68.69%	68.35%	68.35%	68.35%	67.34%	69.70%	69.70%
17	68.69%	68.69%	68.69%	68.69%	68.01%	68.01%	68.01%	65.99%	69.02%	70.71%
19	69.02%	69.02%	69.02%	69.02%	67.68%	67.68%	67.68%	65.66%	67.68%	70.37%
21	68.01%	68.01%	68.01%	68.01%	67.68%	67.68%	67.68%	65.99%	65.32%	70.03%
23	68.01%	68.01%	68.01%	68.01%	67.68%	67.68%	67.68%	67.00%	68.01%	70.37%
25	68.35%	68.35%	68.35%	68.35%	68.01%	68.01%	68.01%	67.34%	68.01%	70.03%
27	68.01%	68.01%	68.01%	68.01%	67.68%	67.68%	67.68%	67.68%	68.01%	69.36%
29	68.01%	68.01%	68.01%	68.01%	68.01%	68.01%	68.01%	68.01%	68.69%	69.70%
31	67.68%	67.68%	67.68%	67.68%	68.01%	68.01%	68.01%	67.68%	68.35%	70.03%
Mean	67.90%	67.90%	67.90%	67.90%	67.21%	67.21%	67.21%	68.26%	68.87%	69.41%
Std. Dev.	0.00897	0.00897	0.00897	0.00897	0.01494	0.01494	0.01494	0.01924	0.01826	0.00917

Another important observation is that the KNN performance was better for the YES class (patients with heart disease) compared to the NO class (patients without heart disease). This suggests that KNN is better at detecting patients with heart disease. Furthermore, the results showed that different correlation filters and evaluation metrics significantly affected KNN performance. The C_F = 1.0 filter and the Cohen's Kappa metric showed the worst results, while the C_F = 0.3 filter and the F-measure metric showed the best results.

Finally, the results show that KNN performance depends on several factors, including the value of K, correlation

Table 5. Comparative analysis of classification performance using KNN algorithm with and without time attribute.

K	Filter	Acuracy	Cohen's Kappa	Class	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure
Best filter combinations and K values with the 'time' attribute													
3	0.3	0.741	0.352	NO	42	24	178	53	0.442	0.636	0.442	0.881	0.522
				YES	178	53	42	24	0.881	0.771	0.881	0.442	0.822
17	1.0	0.731	0.241	NO	22	7	195	73	0.232	0.759	0.232	0.965	0.355
				YES	195	73	22	7	0.965	0.728	0.965	0.232	0.83
17	0.05	0.707	0.172	NO	18	10	192	77	0.189	0.643	0.189	0.95	0.293
				YES	192	77	18	10	0.95	0.714	0.95	0.189	0.815
21	0.10	0.835	0.577	NO	52	6	196	43	0.547	0.897	0.547	0.97	0.68
				YES	196	43	52	6	0.97	0.82	0.97	0.547	0.889
Best filter combinations and K values without the 'time' attribute													
5	0.10	0.731	0.333	NO	42	27	175	53	0.442	0.609	0.442	0.866	0.512
				YES	175	53	42	27	0.866	0.768	0.866	0.442	0.814
9	0.15	0.717	0.245	NO	28	17	185	67	0.295	0.622	0.295	0.916	0.4
				YES	185	67	28	17	0.916	0.734	0.916	0.295	0.815
13	0.5	0.69	0.109	NO	13	10	192	82	0.137	0.565	0.137	0.95	0.22
				YES	192	82	13	10	0.95	0.701	0.95	0.137	0.807
17	0.05	0.707	0.172	NO	18	10	192	77	0.189	0.643	0.189	0.95	0.293
				YES	192	77	18	10	0.95	0.714	0.95	0.189	0.815

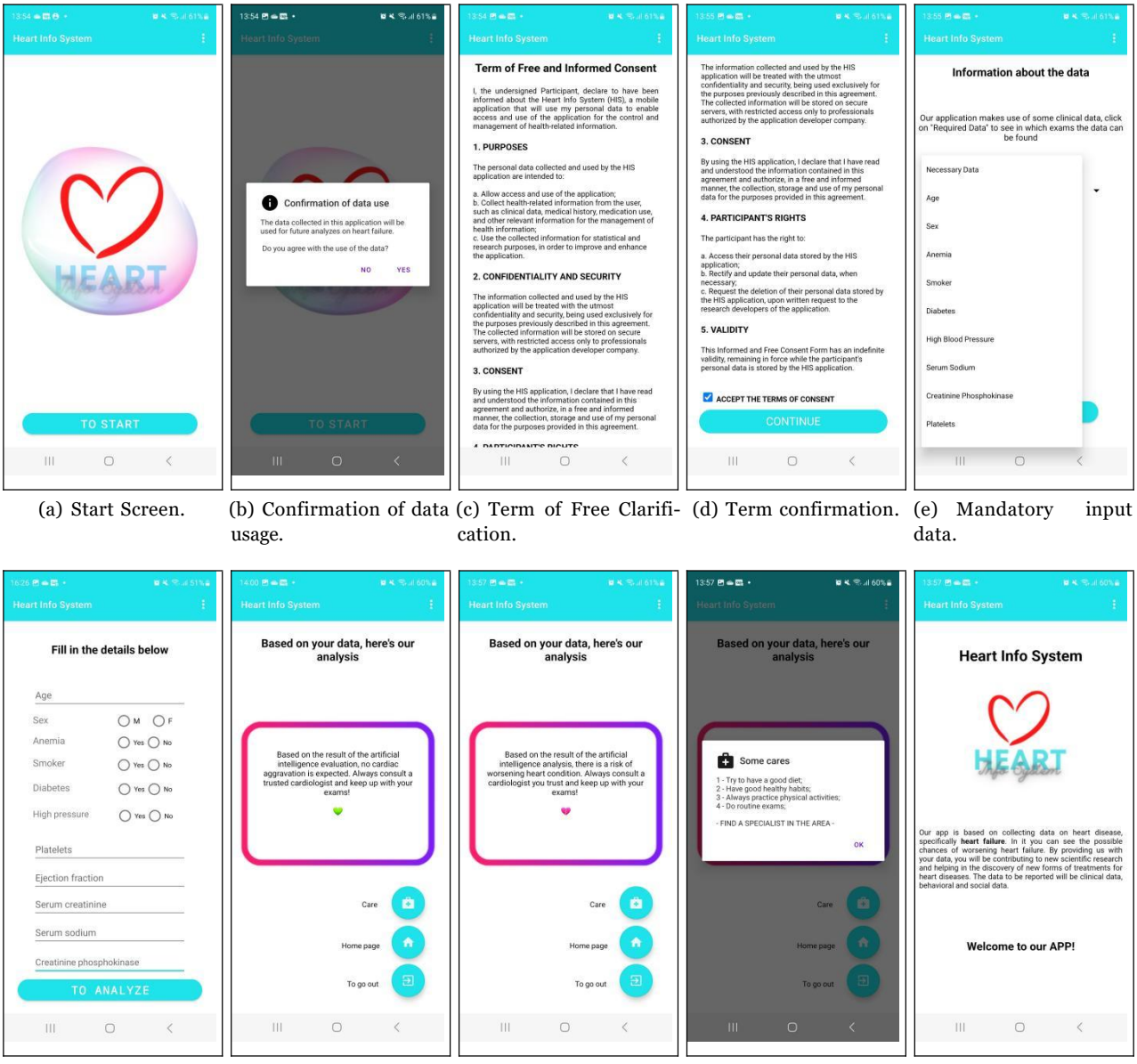
filters and evaluation metrics. Consideration of the time variable may improve KNN performance in detecting heart disease, but more research is needed to determine the best parameter setting for this specific task.

4.3 HIS implementation preview

To assist both physicians and patients, we've created the 'Heart Info System' (HIS) in Java for Android. This application implements an expert system for mobile devices, utilizing the UCI Machine Learning Repository database to diagnose the likelihood of survival in the event of developing heart diseases. Our system is categorized as Artificial Intelligence as a Service (AIaaS), representing a cloud-based service that provides artificial intelligence (AI) outsourcing. AIaaS empowers individuals and businesses to explore and deploy AI for extensive use cases, minimizing risk and eliminating the need for a substantial upfront investment. HIS is a system for mobile devices, on Android,

that uses a KNN algorithm to classify the chance of survival in case of unfolding heart disease. The app allows users to more conveniently monitor their heart health and receive early warnings about potential heart problems. This can lead to better heart health management and reduce the number of heart disease-related deaths.

The Figure 5 presents several screenshots of the Heart Info System (HIS), an Android application that stores patient's information in Firebase database. Each subfigure represents a different screen in the HIS application.



(a) Start Screen. (b) Confirmation of data usage. (c) Term of Free Clarification. (d) Term confirmation. (e) Mandatory input data.

(f) Entry form. (g) Patient without heart problems. (h) Patient with heart problems. (i) Patient's care screen. (j) Screen over the HIS.

Figure 5. Screenshots of different HIS' screens (front-end), an Android application with firebase storage.

Figure 5(a) shows the start screen of the HIS, while Figure 5(b) displays a confirmation of data usage. In Figure 5(c), the term of free clarification is presented, which the user has to accept to proceed. Once the user accepts the terms, the HIS asks for confirmation, as shown in Figure 5(d). Figure 5(e) shows a screen where mandatory input data must be provided. The input form, where the user inputs the patient's data, is presented in Figure 5(f). In Figures 5(g) and 5(h), patients without heart problems and with heart problems, respectively, are presented. Figure 5(i) displays a screen that shows the patient's care, while Figure 5(j) presents a screen over the HIS. These screens are designed to provide a simple and intuitive user experience, while the HIS back-end stores and processes patient's information in a secure manner.

The application consists of a main activity called Analyze, which is responsible for retrieving data from a Firebase database and processing it using the KNN algorithm with (K = 5), value as being the best according to Table 5. Data is retrieved from the Firebase database in two steps. First, data from the original database is retrieved and stored in arrays corresponding to the attributes of the problem. Then, data from the new database are retrieved

and stored in arrays corresponding to the data to be classified. After that, the KNN algorithm is applied to the test data to rank the chance of survival in case of heart disease progression. The application also has some additional features, such as a care button that provides information about care to be taken in case of heart disease and home and exit buttons for user navigation. The availability of this application for mobile devices can make health information more accessible and democratize access to health services.

4.4 Discussions

The analysis carried out showed that the performance of the KNN algorithm is influenced by the values of K and by the correlation filters used. It was observed that increasing the value of K above 21 resulted in a decrease in accuracy, suggesting that the selection of an optimal number of relatively small neighbors is important for model performance. The selection of an adequate correlation filter, with value $C_F = 0.1$, proved to be important to increase the accuracy of the model. The exclusion of the time attribute resulted in a significant loss of information, but it was necessary to avoid possible biases and to generalize the predictions to different clinical scenarios.

Results show that thoughtful correlation filter selection improves survival predictions in heart disease patients using KNN. The accuracy's variability for different filter values highlights the challenge in filter selection, requiring careful consideration during feature selection and modeling. Optimal performance was observed at $C_F = 0.1$ and $K = 21$, achieving 83.50% accuracy. However, it was observed that, for values of K above 21, the accuracy started to decrease, indicating that the selection of an ideal number of relatively small neighbors is important for the performance of the model. The analysis by average accuracy for different values of K for each value of C_F also varied, with the best average accuracy (82.15%) being obtained for the correlation filter $C_F = 0.1$, while the worst performer (67, 21%) was for the correlation filter $C_F = 0.2$.

Another relevant aspect observed in the analysis was the importance of selecting attributes to improve the effectiveness and generalization of the models. In this case, the exclusion of the time attribute resulted in a significant loss of information, and its inclusion may be relevant in specific clinical contexts. However, this time attribute may not be relevant for cases in which the patient is not hospitalized, and the length of hospitalization is a public health policy, which can be different in each country or region , depending on the public resources of each country. As a result, we chose to exclude the time attribute from certain models as it was deemed to be irrelevant in those particular contexts. However, we recognize that in certain clinical contexts, the inclusion of this attribute may be essential for improving the accuracy of the models.

The HIS offers several advantages, including early detection of potential heart problems through the analysis of patient data and the classification of the chance of survival in case of heart disease progression. This early detection can lead to timely treatment and management of heart problems, improving patient outcomes. Additionally, the system's mobile accessibility as an Android app makes it more convenient for users to monitor their heart health on-the-go, democratizing access to health services and information. The use of an expert system with a KNN algorithm ensures accurate diagnoses and predictions of patient outcomes, while the secure storage of patient information in Firebase database ensures privacy and confidentiality. The app's simple and intuitive user interface makes it easy for patients to input their data and receive their results, and the app also includes additional features such as a care button that provides information about care to be taken in case of heart disease. The system is capable of handling large amounts of data, allowing for more accurate analysis of patient data and better predictions of patient outcomes.

However, there are limitations to the Heart Info System, including its limited scope that only focuses on the classification of the chance of survival in case of heart disease progression, and its dependence on the accuracy and completeness of the UCI Machine Learning Repository database. Furthermore, the HIS system does not provide personalized recommendations for heart disease management and care, as it does not take into account individual patient characteristics or medical history.

5 Conclusions and Future Work

This work presents an innovative approach to the diagnosis of heart disease, using the KNN algorithm and different correlation filters to select the most relevant attributes. Results showed that careful selection of correlation filters can lead to better predictions of survival in patients with heart disease using KNN. In addition, feature selection was important to improve the effectiveness and generalization of the models. We hope that these results can contribute to improvements in public health, offering a more accurate and efficient method for the diagnosis of heart diseases, reducing costs for government and discomfort for the patient.

Furthermore, the results suggest that the selection of an ideal number of relatively small neighbors is important for the performance of the model. The analysis by average accuracy for different values of K for each value of C_F also varied, and the best average accuracy was obtained for the correlation filter $C_F = 0.1$ with $K = 5$, disregarding The value of time is used, which is the patient's hospitalization time, since this time is relative and we disregarded it

for an application in which the patient can use the app based on the results of his exams. While excluding the time attribute resulted in some information loss, this step was crucial to prevent potential biases and ensure generalizable predictions across diverse clinical scenarios. Our findings hold significant promise for public health advancements, potentially enabling more accurate and efficient heart disease diagnosis, leading to reduced healthcare costs and improved patient experiences.

Subsequent efforts could rectify limitations by exploring alternative classification algorithms, incorporating more data and attributes for enhanced model accuracy. Additionally, examining diverse K values and C_F correlation filters on different heart disease datasets could gauge the robustness and generalizability of the proposed approach. The HIS Android an AIaaS system, as implemented in this study, showcases the potential of employing machine learning on mobile devices for heart disease diagnosis. The app, utilizing the KNN algorithm, estimates the likelihood of survival in the face of disease progression. It includes features like a care button offering guidance in case of heart disease, catering to both patients and healthcare professionals. Future work may concentrate on refining the app's usability, enhancing the user interface, and integrating it with other digital health tools and technologies. Ultimately, the integration of precise diagnostic methods into mobile devices can significantly impact early detection and treatment, thereby enhancing patients' health and quality of life.

Funding Support

This research was funded by the Foundation for Supporting Research in the state of Minas Gerais (Fapemig).

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in UCI Machine Learning Repository – [Heart failure clinical records Data Set] at DOI: https://doi.org/10.24432/C5Z89R.

References

Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., and Raza, M. A. (2017). Survival analysis of heart failure patients: A case study. *PloS one*, *12*(7), e0181001. https://doi.org/10.1371/journal.pone.0181001

Alvarez, S. A. (2002). An exact analytical relation among recall, precision, and classification accuracy in information retrieval. *Technical Report BCCS-02-01*, 1-22.

Anggoro, D. A. and Kurnia, N. D. (2020). Comparison of accuracy level of support vector machine (svm) and k-nearest neighbors (knn) algorithms in predicting heart disease. *International Journal of Emerging Trends in Engineering Research*, *8*(5), 1689–1694. https://doi.org/10.30534/ijeter/2020/32852020

Brouwers, S., Sudano, I., Kokubo, Y., & Sulaica, E. M. (2021). Arterial hypertension. *The Lancet*, *398*(10296), 249-261. https://doi.org/10.1016/S0140-6736(21)00221-X

Cambronero, C. G. and Moreno, I. G. (2006). Algoritmos de aprendizaje: knn & kmeans. *Intelgencia en Redes de Comunicación, Universidad Carlos III de Madrid*, *23*.

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, *20*, 1-16. https://doi.org/10.1186/s12911-020-1023-5

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more informative than Cohen's Kappa and Brier score in binary classification assessment. *IEEE Access*, *9*, 78368-78381. https://doi.org/10.1109/ACCESS.2021.3084050

de Brito, R. X., Fernandes, C. A. R., & Amora, M. A. B. (2020). Análise de Desempenho com Redes Neurais Artificiais, Arquiteturas MLP e RBF para um Problema de Classificação de Crianças com Autismo. *iSys-Brazilian Journal of Information Systems*, *13*(1), 60-76.

de Oliveira, F. A., dos Santos Villote, G., Costa, R. L., Goldschmidt, R. R., & Cavalcanti, M. C. (2020). Minerando Regras de Associação de Multirrelação na Web de Dados. *iSys-Brazilian Journal of Information Systems, 13*(4), 77-100. https://doi.org/10.5753/isys.2020.830

Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. *Procedia technology*, *10*, 85-94. https://doi.org/10.1016/j.protcy.2013.12.340

Dietz, C., & Berthold, M. R. (2016). KNIME for open-source bioimage analysis: a tutorial. *Focus on Bio-Image Informatics*, 179-197. https://doi.org/10.1007/978-3-319-28549-8_7

Dornelas, R. S., & Lima, D. A. (2023). Correlation filters in machine learning algorithms to select demographic and individual features for Autism Spectrum Disorder diagnosis. *Journal of Data Science and Intelligent Systems*, *1*(2), 105-127. https://doi.org/10.47852/bonviewJDSIS32021027

Enriko, I. K. A., Suryanegara, M., & Gunawan, D. (2018). Heart disease diagnosis system with k-nearest neighbors method using real clinical medical records. In *Proceedings of the 4th international conference on frontiers of educational technologies*, 127-131. https://doi.org/10.1145/3233347.3233386

Fillbrunn, A., Dietz, C., Pfeuffer, J., Rahn, R., Landrum, G. A., & Berthold, M. R. (2017). KNIME for reproducible cross-domain analysis of life science data. *Journal of biotechnology*, *261*, 149-156. https://doi.org/10.1016/j.jbiotec.2017.07.028

Fleury, M. T. L., & da Costa Werlang, S. R. (2016). Pesquisa aplicada: conceitos e abordagens. *Anuário de Pesquisa GVPesquisa*.

Haider, A. W., Larson, M. G., Franklin, S. S., & Levy, D. (2003). Systolic blood pressure, diastolic blood pressure, and pulse pressure as predictors of risk for congestive heart failure in the Framingham Heart Study. *Annals of internal medicine*, *138*(1), 10-16. https://doi.org/10.7326/0003-4819-138-1-200301070-0000

Hu, D., Liang, K., Zhou, S., Tu, W., Liu, M., & Liu, X. (2023). scDFC: A deep fusion clustering method for single-cell RNA-seq data. *Briefings in Bioinformatics*, *24*(4), bbad216. https://doi.org/10.1093/bib/bbad216

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. *IEEE access*, *9*, 39707-39716. https://doi.org/10.1109/ACCESS.2021.3064084

Iwano, H., & Little, W. C. (2013). Heart failure: what does ejection fraction have to do with it?. *Journal of cardiology*, *62*(1), 1-3. https://doi.org/10.1016/j.jjcc.2013.02.017

Japkowicz, N. (2006). Why question machine learning evaluation methods. In *AAAI workshop on evaluation methods for machine learning*, *6*, 11.

Kim, H., Caulfield, L. E., Garcia - Larsen, V., Steffen, L. M., Coresh, J., & Rebholz, C. M. (2019). Plant - based diets are associated with a lower risk of incident cardiovascular disease, cardiovascular disease mortality, and all - cause mortality in a general population of middle - aged adults. *Journal of the American Heart Association*, *8*(16), e012865. https://doi.org/10.1161/JAHA.119.012865

Kumar, D., Verma, C., Dahiya, S., Singh, P. K., Raboaca, M. S., Illés, Z., & Bakariya, B. (2021). Cardiac diagnostic feature and demographic identification (CDF-DI): an IoT enabled healthcare framework using machine learning. *Sensors*, *21*(19), 6584. https://doi.org/10.3390/s21196584

Liang, K., Meng, L., Liu, M., Liu, Y., Tu, W., Wang, S., ... & Sun, F. (2022). A survey of knowledge graph reasoning on graph types: Static, dynamic, and multimodal. *arXiv preprint:2212.05767*.

Lima, D. A., Ferreira, M. E. A., & Silva, A. F. F. (2021). Machine learning and data visualization to evaluate a robotics and programming project targeted for women. *Journal of Intelligent & Robotic Systems*, *103*(1), 4. https://doi.org/10.1007/s10846-021-01443-w

Liu, S., Liu, D., Srivastava, G., Połap, D., & Woźniak, M. (2021). Overview and methods of correlation filter algorithms in object tracking. *Complex & Intelligent Systems*, *7*, 1895-1917. https://doi.org/10.1007/s40747-020-00161-4

Neto, V. V. G., Oquendo, F., & Nakagawa, E. Y. (2017). Smart systems-of-information systems: Foundations and an assessment model for research development. *Sociedade Brasileira de Computação*.

Pérez, J., Díaz, J., Garcia-Martin, J., & Tabuenca, B. (2020). Systematic literature reviews in software engineering—Enhancement of the study selection process using Cohen's kappa statistic. *Journal of Systems and Software*, *168*, 110657. https://doi.org/10.1016/j.jss.2020.110657

Phyu, T. N. (2009, March). Survey of classification techniques in data mining. In *Proceedings of the international multiconference of engineers and computer scientists*, *1*(5), 727-731.

Raikwal, J. S., & Saxena, K. (2012). Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. *International Journal of Computer Applications*, *50*(14).

Raupp, F. M., & Beuren, I. M. (2006). Metodologia da pesquisa aplicável às ciências. *Como elaborar trabalhos monográficos em contabilidade: teoria e prática*, 76-97.

Rodrigues, W. C. (2007). Metodologia científica. *Faetec/IST. Paracambi*, *2*.

Santos, I. D. S., & Bittencourt, M. S. (2008). Insuficiência cardíaca. *Rev. med.(Säo Paulo)*, 224-231.

Siqueira, S., Bittencourt, I., Isotani, S., & Nunes, B. P. (2017). Information systems based on (linked) open data: From openness to innovation. *Sociedade Brasileira de Computação*.

Souza, V. S., & Lima, D. A. (2023). Identifying risk factors for heart failure: A case study employing data mining algorithms. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS32021386

Stamler, J., Neaton, J. D., & Wentworth, D. N. (1989). Blood pressure (systolic and diastolic) and risk of fatal coronary heart disease. *Hypertension, 13*(5_supplement), I2.

Vasan, R. S., Larson, M. G., Leip, E. P., Evans, J. C., O'Donnell, C. J., Kannel, W. B., & Levy, D. (2001). Impact of high-normal blood pressure on the risk of cardiovascular disease. *New England journal of medicine*, *345*(18), 1291-1297.

Warrens, M. J. (2015). Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy*, 5. https://doi.org/10.4172/2161-0487.1000197

Xing, W., & Bei, Y. (2019). Medical health big data classification based on KNN classification algorithm. *IEEE Access*, *8*, 28808-28819. https://doi.org/10.1109/ACCESS.2019.2955754

Yuan, D., Kang, W., & He, Z. (2020). Robust visual tracking with correlation filters and metric learning. *Knowledge-Based Systems*, *195*, 105697. https://doi.org/10.1016/j.knosys.2020.105697

Zhang, S., Cheng, D., Deng, Z., Zong, M., & Deng, X. (2018). A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, *109*, 44-54. https://doi.org/10.1016/j.patrec.2017.09.036

Zhuang, J., Cai, J., Wang, R., Zhang, J., & Zheng, W. S. (2020). Deep kNN for medical image classification. In *Medical Image Computing and Computer Assisted Intervention*, *23*, 127-136. https://doi.org/10.1007/978-3-030-59710-8_13

Stefane Souza, V., & Araújo Lima, D. (2024). Cardiac disease diagnosis using K-Nearest neighbor algorithm: A study on heart failure clinical records dataset. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA42022045