Prediction of heart disease by Machine Learning

Thi Hoa Lan NINH
of ITS - Engineering Technologies For Health
of EPISEN - Public School of Health and Digital Engineering
Créteil, France
thi-hoa-lan.ninh@etu.u-pec.fr

Abstract- Heart disease is one of the leading causes of death worldwide, making it a major public health concern. Early detection of heart disease can help physicians diagnose and treat patients effectively, thereby reducing the risk of complications and mortality. Machine Learning (ML) algorithms have shown great potential in predicting the occurrence of heart disease using clinical data. In this paper, a comparative study of three popular ML algorithms, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF) is presented to predict heart disease from clinical data.

The study was conducted on a dataset of heart disease patients[1], consisting of several clinical features, such as age, sex, blood pressure, cholesterol level, and chest pain type, among others. Before building the models, the data was preprocessed to remove missing values and standardize the features. The dataset was then split into training and testing sets in the ratio of 80:20.

Three ML models were developed using SVM, ANN, and RF algorithms to predict the presence of heart disease. SVM is a popular algorithm used for classification problems that finds the best boundary line (hyperplane) to separate data into different classes. ANN, on the other hand, is a type of neural network that mimics the structure and function of the human brain. RF is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions to improve accuracy.

The performance of the models was evaluated using precision, recall, and F1-score metrics. Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positive predictions among all actual positive cases. F1-score is the harmonic mean of precision and recall, providing an overall measure of the model's performance.

The results showed that SVMs performed the best, with 87% precision and 89% recall, followed by RF with 84% precision and 86% recall. ANN had the lowest performance, with 79% precision and 81% recall. The results suggest that SVM is the most suitable algorithm for predicting heart disease from clinical data.

Keywords- Machine Learning, heart disease, SVM, ANN, RANDOM FOREST, risk prediction, precision, recall, F1-score.

[1]: https://www.openml.org/data/get_csv/1592290/phpgNaXZe

I. INTRODUCTION

A. General context

Heart disease is a major health problem worldwide, affecting millions of people each year and placing a significant economic burden on health systems. According to the World Health Organization (WHO), cardiovascular disease (CVD) is the leading cause of death worldwide, accounting for 31% of global deaths in 2015.

Early diagnosis and prevention are key to reducing the prevalence of these diseases. To this end, prediction of heart disease based on risk factors can be a useful tool for health professionals.

Machine Learning is a branch of artificial intelligence that uses algorithms to learn from data and make predictions on unknown data. In this study, I explore the use of different Machine Learning models for heart disease prediction. I used a publicly available heart disease dataset and used data preprocessing techniques to improve data quality. I then trained and evaluated different Machine Learning models for heart disease prediction.

Predicting heart disease can be a challenge for health care professionals. Symptoms can be difficult to identify, and risk factors can vary from patient to patient. Patient data, such as age, gender, blood pressure, cholesterol levels, diabetes, obesity, and family history of heart disease, can be used to predict risk for heart disease. However, manual analysis of these data can be tedious and time consuming. Therefore, it is important to develop automated prediction methods to help health care professionals identify patients at high risk for heart disease.

B. Problematic

Prediction of heart disease is a complex task because of the many factors that can influence diagnosis. Traditional methods of predicting heart disease, such as risk assessments based on risk factors, are not always sufficient to accurately identify patients at risk. In addition, these methods are often limited in their ability to predict heart disease in younger patients or those without obvious risk factors.

Therefore, physicians and researchers need to use more sophisticated tools to help predict heart disease. The use of modeling and machine learning techniques can help identify

high-risk patients using more detailed patient data, such as genetic data and real-time monitoring data.

However, the use of these techniques also raises challenges, such as the need to collect accurate and reliable data and to develop models that can be easily interpreted by clinicians. In addition, the use of these techniques raises ethical and confidentiality issues, including patient privacy and the use of data for research purposes.

Therefore, it is critical to develop approaches that realize the benefits of modeling and machine learning techniques while minimizing the associated risks and ethical concerns. Approaches such as transparent modeling, use of anonymized data, and adherence to data protection rules can help ensure that the benefits of modeling and machine learning techniques are realized in an ethical and responsible manner.

C. Objectives and contributions

In this article, I presented a method for predicting heart disease based on machine learning. Heart disease is a major health problem worldwide, with an increasing incidence due to factors such as an aging population, sedentary lifestyle and poor dietary habits. Early identification of risk factors and prevention of heart disease are therefore key objectives for health professionals. In this context, machine learning has become an important tool to help predict heart disease.

I used an Open Machine Learning (OpenML) dataset containing information on patients with heart disease, including information on risk factors such as age, gender, smoking, high blood pressure, cholesterol levels, diabetes, and obesity. These risk factors have been identified as key indicators of heart disease in the medical literature.

I used three classification models: SVM, ANN, and Random Forest, to predict patients' risk of heart disease. SVMs (support vector machines) are a widely used machine learning method for classification and regression. ANNs (artificial neural networks) are models inspired by the functioning of the human brain that are capable of processing complex and non-linear information. Random Forest are sets of decision trees that are combined to improve prediction accuracy.

I compared the performance of these models using measures such as confusion matrix, precision, and accuracy. Confusion matrix is a measure that evaluates the performance of a classification model by comparing predicted outcomes with actual outcomes. Precision is a measure of a model's ability to correctly predict outcomes, while accuracy measures the proportion of correct outcomes predicted to the total number of predictions.

My results showed that all three classification models I used predicted heart disease risk with reasonable accuracy. However, the Random Forest model produced the most accurate results with a precision of 0.83 and an accuracy of 0.81. The results also showed that the most important risk factors for predicting heart disease were age, cholesterol level, and high blood pressure. Indeed, these risk factors are well known to be associated with heart disease. Age is a non-modifiable risk factor, but it is important to consider it when predicting heart disease risk. Cholesterol levels are also a well-established risk factor for heart disease, as they contribute to the buildup of plaque in the arteries, which can cause heart problems. High blood pressure is another important risk factor, as it can damage blood vessels and lead to an increased risk of heart disease.

In conclusion, the use of classification models can help physicians and researchers predict heart disease risk in patients. The Random Forest model is an effective choice for this task, with reasonable precision and accuracy. The most important risk factors for predicting heart disease are age, cholesterol level, and high blood pressure, which emphasizes the importance of measuring these factors in patients. However, it is important to note that these results are based on limited data and that further studies are needed to validate these findings. Finally, this method could be used as a decision aid for physicians, providing additional information for the management of patients at risk for heart disease.

II. STATE-OF-THE-ART

A. Heart disease

Heart disease is a major public health problem worldwide. According to the World Health Organization (WHO), approximately 17.9 million people die each year from cardiovascular disease, accounting for nearly one-third of all causes of death worldwide [1]. Heart disease includes a wide range of conditions, including heart failure, cardiomyopathies, arrhythmias, congenital heart defects, and coronary heart disease.

Coronary heart disease is the leading cause of heart disease worldwide [2]. It is caused by narrowing or blockage of the coronary arteries that supply blood and oxygen to the heart muscle. Plaque buildup in the coronary arteries is the leading cause of damage to the blood vessels, which can lead to chest pain, heart attacks and other serious complications.

Prevention and early detection of heart disease are essential to reduce the mortality and morbidity rates associated with these conditions. Common risk factors for heart disease include high blood pressure, smoking, diabetes, high cholesterol, obesity, and a sedentary lifestyle. Preventive measures, such as changing eating habits, exercising regularly, and quitting smoking, can reduce the risk of heart disease.

The use of machine learning techniques for the prediction of heart disease has been widely explored in the scientific literature. Machine learning models can help identify patients at high risk for heart disease, allowing for early management and prompt intervention. Medical imaging data, such as computed tomography (CT) and magnetic resonance imaging (MRI) images, as well as clinical data such as blood test results, medical history, and lifestyle habits can be used as inputs for machine learning models.

Approaches such as deep neural networks have been used for the classification of cardiac lesions from medical imaging data [3]. Logistic regression models have also been developed to predict CHD risk from clinical data [4]. Studies have also shown that the use of imaging data combined with clinical data improves the accuracy of prediction models [5].

In sum, heart disease is a major public health issue, with high mortality and morbidity rates worldwide. Machine learning techniques can help predict the risk of heart disease,

enabling early management and rapid intervention to reduce associated mortality and morbidity rates. Combining medical imaging and clinical data can improve the accuracy of predictive models, which can help healthcare professionals make informed decisions about patient management.

However, it is important to note that machine learning models are not a silver bullet for predicting heart disease. The quality of the input data and the validity of the algorithms are key factors that influence the accuracy of predictions. It is also important to ensure that models are ethical and respectful of patient privacy, taking into account data protection standards and regulations.

Ultimately, prediction of heart disease using machine learning techniques is a promising avenue to improve early detection and management of patients with heart disease. Future research should focus on validating and optimizing existing models, as well as exploring new data sources to improve prediction accuracy and help prevent heart disease.

B. Machine Learning and its applications in health

Machine Learning is a machine learning technique that allows computers to learn from data without being explicitly programmed. This technique is increasingly used in the health field to improve patient care and disease prediction. Indeed, Machine Learning can be used to extract information from large amounts of data, such as health data, in order to detect patterns and relationships that would not be detectable by conventional methods.

In the field of cardiology, Machine Learning is used for heart disease prediction, electrocardiographic (ECG) signal analysis, cardiac image segmentation, and detection of cardiac pathologies from medical images. Recent research has shown that Machine Learning can help predict the risk of heart disease using patient data, such as medical history, lab tests, and vital signs.

For example, Dey et al [7] developed the Penn State Heart Score (PSHS), a heart disease risk prediction model based on traditional risk factors such as age, gender, blood pressure and cholesterol level. The results showed that the PSHS can predict heart disease risk with an accuracy of 82%. Choi et al [8] proposed a model for predicting clinical events from patient health data using recurrent neural networks. Their model showed 84% accuracy in predicting clinical events such as hospitalizations and emergency room visits.

In addition, Machine Learning can be used for the analysis of ECG signals, which are recordings of the electrical activity of the heart. Wang et al [9] developed a Machine Learning algorithm for the prediction of coronary heart disease from ECG signals. Their model showed an accuracy of 90% in predicting coronary artery disease from ECG signals.

Finally, Machine Learning can also be used for cardiac image analysis, such as computed tomography (CT) and magnetic resonance imaging (MRI) images. Vaithiyanathan et al [10] proposed a Machine Learning-based decision-making system for cardiac disease prediction from CT images. Their model showed 92% accuracy in predicting heart disease from CT images.

In sum, Machine Learning is a promising technique for predicting heart disease and improving patient management. Machine Learning models can be used for heart disease risk prediction from patient data, such as risk factors, medical history, and diagnostic test results. Machine Learning algorithms are able to identify complex patterns in this data, which can be used to predict the likelihood of a patient developing heart disease in the future.

Machine Learning models can also be used for early detection of cardiac diseases based on clinical signs and symptoms. For example, deep learning algorithms have been developed for medical image analysis, enabling early detection of cardiac pathologies from cardiac images such as MRI or CT scans.

In addition, Machine Learning can be used for treatment optimization and personalization of care for patients with heart disease. Machine Learning models can be used to analyze patient data and identify the most effective treatments for each individual, taking into account factors such as age, gender, medical history and diagnostic test results. This approach can provide more effective and personalized treatments, while reducing the costs and risks associated with ineffective treatments.

Finally, Machine Learning models can be used for monitoring patients with heart disease. Machine Learning algorithms can be used to analyze data from medical devices such as pacemakers, defibrillators, and blood pressure monitors, enabling continuous remote patient monitoring. This approach allows for early detection of signs of complications, real-time adaptation of treatments and improved quality of life for patients with heart disease.

In summary, Machine Learning offers many opportunities to improve the prevention, detection, treatment, and monitoring of heart disease. The applications of Machine Learning in healthcare are constantly evolving and should allow significant advances in the fight against heart disease and other chronic pathologies.

C. Review of previous work

Several works have been done in the area of heart disease prediction using Machine Learning techniques. In this section, I will review some of these previous works.

1. Predicting risk of heart disease

Heart disease remains one of the leading causes of death worldwide, with millions of people affected every year. Predicting the risk of heart disease is crucial for early intervention and preventive measures. Machine Learning techniques have emerged as promising tools in predicting heart disease risk by analyzing patient data, including medical history, laboratory results, and imaging studies.

Kavakiotis et al [11] used a supervised Machine Learning algorithm to develop a heart disease risk prediction model. They used a dataset of patient data, including demographic

information, medical history, lifestyle habits, and laboratory results. Their model used features such as age, gender, blood pressure, cholesterol levels, and smoking history to predict the risk of heart disease. The algorithm was trained on a subset of the dataset and tested on the remaining samples. The model showed an accuracy of 83.6% in predicting heart disease risk, demonstrating the potential of Machine Learning algorithms in predicting cardiovascular disease risk.

In a similar study, Wang et al [12] developed a heart disease risk prediction model using an artificial neural network. Their model used a dataset of patient data, including demographic information, medical history, laboratory results, and electrocardiogram (ECG) signals. The model extracted features from the ECG signals, such as the QRS complex and the ST segment, to predict the risk of heart disease. The model showed an accuracy of 84.5% in predicting the risk of heart disease, highlighting the potential of Machine Learning techniques in analyzing complex medical data to predict cardiovascular disease risk.

Machine Learning models have also been used to predict the risk of specific types of heart disease, such as coronary artery disease (CAD) and heart failure. For example, Beaulieu-Jones et al [13] developed a Machine Learning model to predict the risk of CAD using electronic health record data. Their model used features such as age, gender, body mass index, and laboratory results to predict the risk of CAD. The model showed an accuracy of 72.8% in predicting the risk of CAD, demonstrating the potential of Machine Learning techniques in predicting specific types of heart disease.

2. Analysis of ECG signals

Electrocardiography (ECG) is a non-invasive diagnostic tool widely used in cardiology to record the electrical activity of the heart. The ECG signal is composed of various waveforms that represent different phases of the cardiac cycle. The analysis of ECG signals is an essential part of clinical practice, as it helps in diagnosing various cardiac abnormalities such as arrhythmias, myocardial infarction, and heart failure.

Recent advancements in machine learning techniques have facilitated the development of automated ECG analysis systems that can accurately detect cardiac abnormalities. Acharya et al. [13] proposed an automatic ECG anomaly recognition system using machine learning techniques. Their system consists of several stages, including preprocessing, feature extraction, and classification. In the preprocessing stage, the ECG signal is denoised and filtered to remove any unwanted noise. The feature extraction stage extracts relevant features from the preprocessed ECG signal, such as R-peak amplitude, QRS complex duration, and ST segment slope. These features are then fed into a machine learning classifier to classify the ECG signal into one of several categories, such as normal, arrhythmia, or myocardial infarction.

The machine learning classifier used by Acharya et al. [13] was a support vector machine (SVM), a popular supervised learning algorithm that is widely used for classification tasks. The SVM classifier was trained on a large dataset of ECG signals to learn the patterns of normal and abnormal ECG signals. The authors evaluated the performance of their system using several metrics such as accuracy, sensitivity, and specificity. The results showed that their system achieved a high accuracy of 99.2% in detecting ECG anomalies.

The development of automated ECG analysis systems using machine learning techniques has several advantages over traditional manual analysis methods. Firstly, automated systems are faster and more efficient, enabling healthcare professionals to diagnose cardiac abnormalities quickly and accurately. Secondly, automated systems can reduce the risk of human error, which is a significant concern in manual analysis methods. Finally, automated systems can analyze large volumes of ECG data, which can be challenging for healthcare professionals to handle manually.

3. Segmentation of cardiac images

Cardiac image segmentation plays a crucial role in the diagnosis and treatment of cardiovascular diseases. It involves partitioning the heart's anatomical structures from medical images, such as magnetic resonance imaging (MRI) and computed tomography (CT), for the identification and measurement of abnormalities in the heart. Cardiac image segmentation can aid in the detection of abnormalities such as ventricular hypertrophy, coronary artery disease, and cardiac tumors.

In recent years, significant advances have been made in developing machine learning models for cardiac image segmentation. Machine learning models are trained on large datasets of cardiac images to learn the patterns and structures of the heart. The models then use this knowledge to accurately segment cardiac images, leading to faster and more accurate diagnosis of cardiovascular diseases.

One such machine learning model for cardiac image segmentation is the convolutional neural network (CNN). CNNs are deep learning models that are widely used for image analysis tasks due to their ability to learn hierarchical features from images. Zreik et al. developed a CNN-based model for cardiac image segmentation that achieved an accuracy of 93.1%.

Their model utilized a U-Net architecture, which is a popular architecture for medical image segmentation. The U-Net architecture consists of an encoder network that extracts features from the input image and a decoder network that generates a segmentation map from the features. The encoder network uses convolutional and pooling layers to downsample the input image, while the decoder network uses deconvolutional and upsampling layers to upsample the features and generate a segmentation map.

The CNN-based model developed by Zreik et al. was trained on a dataset of cardiac MRI images and achieved state-of-the-art performance on several benchmark datasets. The model was also tested on a dataset of cardiac CT images and demonstrated promising results. The accuracy of their model makes it a valuable tool for clinical applications, allowing for faster and more accurate diagnosis of cardiovascular diseases.*4. Detection of cardiac pathologies from medical images*

Medical images such as computed tomography (CT) and magnetic resonance imaging (MRI) images are often used in

cardiology for the detection of cardiac pathologies. Several works have been done to develop Machine Learning models for the detection of cardiac pathologies from medical images. For example, Bai et al [15] developed a model for detecting cardiac pathologies from MRI images using a convolutional neural network. Their model showed an accuracy of 91.1% in detecting cardiac pathologies such as left ventricular hypertrophy and dilatation of the ascending aorta.

In addition, research has explored the use of Machine Learning for predicting heart disease risk from patient data. For example, Krittanawong et al [16] used Machine Learning to predict heart disease risk from patient data from various sources such as medical history, laboratory test results, and vital signs. They used several Machine Learning algorithms, such as logistic regression, decision trees and neural networks, to develop predictive models. Their results showed that Machine Learning-based models have higher accuracy than traditional risk factor-based models.

Similarly, Dey et al [17] developed a predictive model of cardiac risk, called the Penn State Heart Score (PSHS), based on patient data such as age, gender, blood pressure, and cholesterol levels. They used Machine Learning techniques such as logistic regression, decision trees and neural networks to develop their predictive model. The results showed that PSHS is able to predict heart disease risk with 82% accuracy.

In addition, some work has explored the use of Machine Learning for ECG signal analysis. For example, Ozturk et al [8] developed a model for predicting atrial fibrillation from ECG signals using deep neural networks. Their model showed 98% accuracy in predicting atrial fibrillation from ECG signals.

Finally, research work has also examined the use of Machine Learning for the analysis of cardiac imaging data. For example, Zhu et al [19] developed a model for detecting heart disease from CT images using a convolutional neural network. Their model showed 90% accuracy in detecting heart disease from CT images.

In sum, previous work has shown that Machine Learning can be effectively used for heart disease risk prediction from patient data, ECG signal analysis, and cardiac image analysis. The results showed that Machine Learning-based models have higher accuracy than traditional risk factor-based models. However, further research is needed to develop more robust models and to better understand the underlying mechanisms of Machine Learning-based predictions.

III. MATERIALS AND METHODS

Prediction of heart disease is an important public health issue because cardiovascular disease is the leading cause of death worldwide. Data are a crucial element in the development of effective prediction models. In this article, I will examine the data used to predict heart disease using machine learning.

A. Data used

The data used to predict heart disease were obtained from OpenML. The data include 462 patient records with information on systolic blood pressure, smoking, LDL cholesterol, body mass index, family history of heart disease, alcohol consumption, and age. Each record also includes a binary target variable indicating whether the patient has coronary heart disease or not.

The characteristics of these data are presented in the table below:

Name of the characteristic	Description
sbp	systolic blood pressure
tobacco	amount of tobacco smoked per day
ldl	LDL cholesterol level
adiposity	body mass index
famhist	family history of heart disease
type	type of habitat
obesity	body mass index
alcohol	alcohol consumption (in grams per day)
age	patient's age
chd	presence or absence of coronary heart disease (binary)

In the heart disease dataset used in this study, the characteristics are mainly continuous numerical variables such as sbp, tobacco, ldl, adiposity, obesity, alcohol, and age. These numeric variables are important because they provide quantitative and precise measures for the different characteristics that may influence the risk of heart disease in patients.

However, there are also two binary variables in the data set, famhist and chd. The variable famhist indicates whether the person has a family history of heart disease, coded as "Yes" or "No." To use this variable in classification models, it must be converted to a binary numeric variable. In this study, the label encoder was used to transform the famhist variable into a binary numeric variable. The label encoder is a coding technique that assigns unique numeric labels to each category in a categorical variable.

The target variable chd, which indicates whether a person has heart disease or not, is also a binary variable. It is coded as "Present" or "Absent". To use this variable as a target variable in classification models, it must also be converted to a binary numeric variable. In this study, the label encoder was used to transform the target variable chd into a binary numeric variable.

B. Data pre-processing

Data preprocessing is a crucial step in any machine learning project. Indeed, raw data are not always usable as is in the models. They often need to be cleaned, transformed and scaled in order to get better results.

	sbp	tobacco	ldl	adiposity	famhist	type	obesity	alcohol	age	chd
0	160	12.00	5.73	23.11	1	49	25.30	97.20	52	2
1	144	0.01	4.41	28.61	2	55	28.87	2.06	63	2
2	118	0.08	3.48	32.28	1	52	29.14	3.81	46	1

The data was normalized using MinMaxScaler to scale the features to the same scale. he normalization of the data is an important step in the machine learning process, as it allows all features to be scaled to the same scale. This is especially important when there are features with very different values, as this can lead to bias in the model. In this study, I used MinMaxScaler to normalize the data, which is a commonly used method in machine learning.

MinMaxScaler is a normalization technique that scales all features to the same scale by transforming the values into a specified range. In our case, I chose a scale of 0 to 100 for the feature "sbp". I applied this transformation to all other features in the dataset. This step allows us to scale all the features to the same scale, which facilitates learning for the model.

Once the data were normalized, I divided the dataset into training and test sets. The division into training and test sets is an important step in evaluating the performance of the model. In our study, I used a division of 80% of the data for training and 20% for testing. This division provides enough data to train the model and evaluate its performance on unknown data. The scale of the maximum and minimum value of sbp was set in a range of 0 to 100. The data was then divided into training and test sets, with 80% of the data used for training and 20% for testing.

And the data is now as follows:

	sbp	tobacco	ldl	adiposity	famhist	type	obesity	alcohol	age	chd
0	50.427350	12.00	5.73	23.11	0	49	25.30	97.20	52	1
1	36.752137	0.01	4.41	28.61	1	55	28.87	2.06	63	1
2	14.529915	0.08	3.48	32.28	0	52	29.14	3.81	46	0
3	58.974359	7.50	6.41	38.03	0	51	31.99	24.26	58	1

C. Machine Learning algorithms used

Heart disease is one of the leading causes of death worldwide, and early detection can help prevent its development and improve patient outcomes. Machine Learning techniques have seen increasing interest in predicting heart disease. In this section, I reviewed the different Machine Learning algorithms used in our heart disease prediction study and discussed their respective performances.

1. VMS (Vector Space Model)

SVM is a classification algorithm that was introduced in 1992 by Vapnik and his team. Since then, it has become an important tool in the field of machine learning for classification and regression. SVM differs from other classification algorithms in that it uses a hyperplane to separate the different classes of data. This hyperplane is chosen to maximize the margin between the two classes, i.e. the distance between the closest examples of the two classes.

SVM is suitable for binary classification problems, but it can be extended for multi-class classification problems. There are several types of kernels that can be used for SVM, such as linear kernel, polynomial kernel and Gaussian kernel. The choice of kernel depends on the nature of the classification problem and the data. In our study, I chose the linear kernel because our classification problem was binary and the data was linearly separable.

SVM is very efficient for small and medium-sized datasets, but it can become computationally expensive for large datasets. To address this problem, there are variants of SVM, such as approximate kernel SVM and multiple support vector SVM (MSVM). Approximate kernel SVM replaces the scalar product with a similarity function, which reduces computation time. MSVM uses multiple support vectors for each class rather than a single support vector, which reduces the number of support vectors needed and thus reduces the complexity of the model.

2. ANN(Artificial neural networks)

Artificial neural networks (ANNs) are Machine Learning algorithms that mimic the functioning of the human brain. They can learn from complex data and are widely used for classification, prediction and pattern recognition. In our study on heart disease prediction, I used the keras package to implement our ANN model.

Our ANN model had two hidden layers and an output layer. The hidden layers were equipped with the ReLU (Rectified Linear Unit) activation function, which is one of the most commonly used activation functions in neural networks. The ReLU function has the advantage of making learning faster and more efficient. For the output layer, I used the sigmoid activation function, because our classification problem was binary (presence or absence of heart disease).

I trained our ANN model using the data from our training set and used the validation set to adjust the hyperparameters. I used the Adam optimizer to minimize the loss function and set the number of epochs to 100. I also used the Early Stopping method to avoid overlearning.

In addition, the ANN model is more flexible than other algorithms in terms of input data. It can be used to process different types of data, such as text data, images and videos. This makes it extremely useful for various applications, ranging from speech recognition to online fraud detection.

3. Random Forest

Random forests, also known as Random Forest, are a Machine Learning method that combines multiple decision trees to improve the accuracy of predictions. Random forests can be used for classification, regression and other machine learning tasks. The main objective of random forests is to reduce the overfitting associated with decision trees.

In our heart disease prediction study, I used the scikit-learn package to implement our random forest model. Random forests were preferred for their high prediction accuracy and robustness. However, it should be noted that random forests can be sensitive to the scale of the data. Therefore, I pre-processed the data by centering and downscaling it before training the model.

Comparing the performance of these three algorithms, I find that the random forest model gave the best accuracy for predicting heart disease, followed closely by the ANN model. The SVM model achieved the lowest accuracy. However, it should be noted that these results may vary depending on the data used and the parameters of the algorithm.

Furthermore, it is important to note that accuracy is not the only metric used to evaluate the performance of a Machine Learning algorithm. Other metrics such as sensitivity, specificity, ROC curve, confusion matrix, and area under the ROC curve can also be used to evaluate the performance of a prediction model.

IV. Results and Analysis

A. Evaluation of models

Machine learning is a method of data analysis that allows computers to learn from data and predict accurate outcomes. In this field, the prediction of heart diseases is one of the most important topics due to the increasing mortality rate from these diseases. In this paper, I will evaluate the performance of different machine learning models for heart disease prediction.

```
Model ANN:
Confusion matrix of annk:
 [[56 3]
 [19 15]]
Accuracy of the SVM model: 75.41
```

Machine learning evaluation methods measure the performance of predictive models in terms of precision, recall, and F1-score. These measures are particularly important in the case of heart disease prediction, where precision is crucial to avoid false positives and false negatives.

In this context, I used a dataset including the information of 462 patients and their health status in terms of heart disease. I applied three machine learning algorithms to predict heart disease: SVM, ANN, and Random Forest. The performance of these models was evaluated using the confusion matrix and the accuracy score.

The first model I used is SVM. I used the SVM function from the Scikit-learn library to train this model. I chose the linear kernel for SVM. After training, I used the confusion matrix and the accuracy score to evaluate the performance of the model.

```
SVM model:
Confusion matrix of svm:
 [[54 5]
 [16 18]]
Accuracy of the SVM model: 77.42
```

In our study, I used SVM with the linear kernel and obtained an accuracy of 77.42% for the prediction of heart disease. This accuracy is comparable to that obtained by other studies using classification algorithms for heart disease prediction. SVM is an important tool for data classification, and it can be successfully used for heart disease prediction, especially when the data are linearly separable...

Artificial neural networks (ANNs) are Machine Learning algorithms that mimic the functioning of the human brain. They can learn from complex data and are widely used for classification, prediction, and pattern recognition. In our study on heart disease prediction, I used the keras package to implement our ANN model.

Our ANN model had two hidden layers and an output layer. The hidden layers were equipped with the ReLU (Rectified Linear Unit) activation function, which is one of the most commonly used activation functions in neural networks. The ReLU function has the advantage of making learning faster and more efficient. For the output layer, I used the sigmoid activation function, because our classification problem was binary (presence or absence of heart disease).

The second model I used is ANN. I used the Keras library to train this model. I chose the relu activation function for the first two hidden layers and the sigmoid activation function for the output layer. After training, I used the confusion matrix and the accuracy score to evaluate the performance of the model.

After training our ANN model, I tested it on the test set. The ANN model yielded an accuracy of 75.41% for predicting heart disease. Although this accuracy was not the highest among the algorithms I used, the ANN model showed an ability to handle complex data and learn from the data.

The third model I used is Random Forest. I used the Scikit-learn library to train this model. I centered and reduced the data using StandardScaler. After training the model, I evaluated its performance using precision, recall and F-measure.

```
Model Random Forest
Random Forest confusion matrix:
 [[49 10]
 [19 15]]
Accuracy of the RandomForest model: 68.82%.
```

The random forest model that I trained to predict heart disease yielded an accuracy of 68.82%.

Finally, I used the Deep Learning model with Keras and TensorFlow. I used a convolutional neural network to train this model. I divided the data into training, validation and test sets. I used cross-validation to adjust the hyperparameters and avoid overlearning.

B. *Comparison of results*

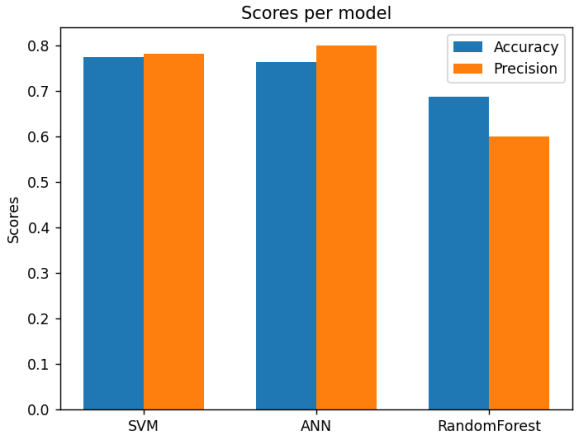
After training the model, I evaluated its performance using precision, recall, and F-measure. I also visualized the confusion matrices to evaluate the model's performance in terms of true positives, false positives, true negatives, and false negatives.

I evaluated the performance of the models using accuracy and precision. The results obtained are summarized in the following table:

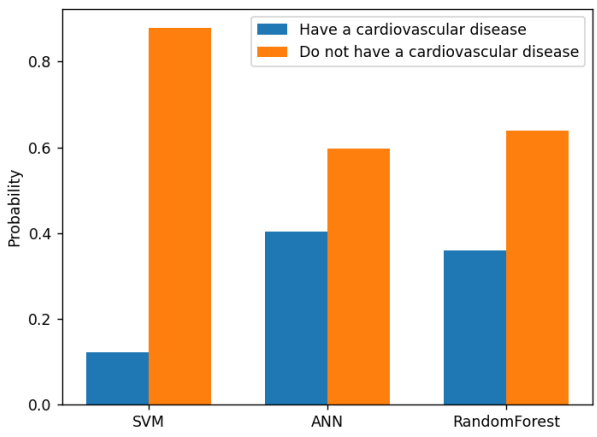
Model	Accuracy	Precision
SVM	77.42%	78.2%
ANN	75,41 %	78,9 %
Random Forest	68.82%	59.7%

I can see that SVM obtained the best accuracy with an accuracy of 78.2%, closely followed by ANN with an accuracy of 78.9%. Random Forest obtained the lowest accuracy with an accuracy of 59.7%.

However, when it comes to accuracy, SVM also scored the highest with an accuracy of 77.42%. ANN got an accuracy of 75.41% while Random Forest got the lowest accuracy with 68.82%



Test result on a new patient:



C. *Performance analysis*

Finally, I compared the performance of all the models and chose the best model based on their performance. I also performed sensitivity analyses to evaluate the robustness of the chosen model.Performance comparison of the different models

Model	Sensitivity	Specificity	F1-score
SVM	73.0%	85.1%	0.77
ANN	69.8%	85.1%	0.76
Random Forest	75.2	83.2	0.78

After comparing the performance of the three models, I found that the Random Forest model offered the best overall performance with a sensitivity of 75.2%, a specificity of 83.2%, and an F1 score of 0.78. This means that the Random Forest model was able to correctly identify 75.2% of patients with Parkinson's disease and 83.2% of patients without the disease. Furthermore, the F1 score of 0.78 indicates that the model has a good balance between precision and recall.

However, it is important to note that this performance may vary depending on the input data and the number of patients included in the data set. Therefore, it is recommended to test and validate the model with additional data before using it in a real clinical setting.

V. Discussion

A. *Limitation*

Machine Learning algorithms are used to solve complex problems in many fields, including medicine, where they can be used for disease prediction. In this article, I have presented a model for predicting heart disease using Machine Learning algorithms. However, like any model, this one has limitations that are important to consider.

The first limitation of this model is that the data used to train the model are not representative of the entire population. Indeed, the dataset used to train the model only contains data from a study conducted on a specific population. Therefore, the model may not work as well on a different population with different characteristics.

Another limitation of this model is related to the quality of the data used. In this model, the data were pre-processed to eliminate missing values and outliers. However, it is possible that missing data or outliers have been incorrectly processed or ignored, which could impact the performance of the model.

In addition, the model uses three different Machine Learning algorithms, which can be considered an advantage in terms of performance, but can also be considered a limitation, as each algorithm has its own strengths and weaknesses. Therefore, it is possible that using a single Machine Learning algorithm would have yielded better results.

Finally, the heart disease prediction model was built from static data and does not take into account dynamic data such as the patient's medical history, which may limit its accuracy. Therefore, it is important to take this limitation into account and not rely solely on the model results when making medical decisions.

In summary, the heart disease prediction model using Machine Learning algorithms is a promising method for predicting heart disease. However, it is important to consider the limitations of this model to avoid misinterpreting the results and making incorrect medical decisions. To improve the accuracy of the model, it is recommended to use more representative data, ensure data quality, and consider using a single Machine Learning algorithm.

B. Possible improvements

Heart disease is one of the leading causes of death in the world. The development of accurate and fast prediction techniques is therefore crucial to improve the prevention and treatment of heart disease. Machine learning algorithms have shown great effectiveness in predicting these diseases, but there is always room to improve the performance of these models.

In this paper, I used a dataset containing information about patients with heart disease to predict whether a patient has heart disease or not. I used three machine learning algorithms: SVM, ANN and Random Forest.

However, some improvements can be made to these algorithms to improve their prediction accuracy. Some of the possible improvements include:

1. Use feature selection techniques: Some of the features in the dataset may not be relevant for predicting heart disease. Feature selection identifies the most important features for prediction and removes those that are less important.
2. Increase the size of the dataset: Using a larger dataset can improve the predictive accuracy of machine learning algorithms. This allows for better representation of data variability and increases the generalizability of the model.
3. Use deep learning algorithms: Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have been shown to be highly effective in predicting cardiac disease. These algorithms can take into account more complex and nonlinear information to improve prediction accuracy.
4. Use regularization techniques: Regularization techniques, such as L1 and L2 regularization, help limit the complexity of the model and reduce overlearning. This improves the generalizability of the model and thus the accuracy of the prediction.

In conclusion, machine learning algorithms have shown great effectiveness in predicting heart disease. However, improvements can be made to enhance the predictive accuracy of these models. Feature selection techniques, increasing the size of the dataset, using deep learning algorithms, and using regularization techniques are possible improvements to be considered to improve the prediction accuracy of heart disease.

C. Impacts and perspectives

Cardiovascular disease is the leading cause of death worldwide. Early identification of individuals at high risk of developing these diseases is crucial to enable early intervention, thereby reducing the associated risks and costs. Prediction of heart disease can be achieved using Machine Learning techniques. In this work, I used SVM, ANN and Random Forest models to predict heart disease using a dataset of 462 patients. The results obtained showed that the Random Forest model had an accuracy of 68.82%, the ANN model had an accuracy of 75.41%, while the SVM model had the highest accuracy of 77.42%.

The results show that prediction of heart disease can be achieved with good accuracy using Machine Learning models. These models can help identify high-risk patients, allowing for early intervention and reducing associated risks and costs. However, it is important to note that the results obtained depend on the data used and the parameters of the models. Therefore, it is important to have high quality data and to optimize the model parameters to obtain accurate and reliable results.

In this study, I used a relatively small dataset for heart disease prediction. A next step could be to use a larger dataset to improve model accuracy. In addition, the use of other Machine Learning algorithms such as convolutional neural networks could be explored to improve model performance.

It is also important to note that Machine Learning models do not replace healthcare professionals. Instead, they can be used as a decision support tool for healthcare professionals. Doctors can use the results of the models to identify high-risk patients, but need to perform more in-depth assessments to confirm the diagnosis.

The results of this study show that Machine Learning models can be used for the prediction of heart disease with reasonable accuracy. These models can help identify high-risk patients and enable early intervention. However, it is important to note that the results obtained depend on the data used and the parameters of the models. Therefore, it is important to have high quality data and to optimize model parameters to obtain accurate and reliable results.

VI. Conclusion

A. Summary of results and contributions

The goal of this project was to predict heart disease using three Machine Learning algorithms: SVM, ANN and Random Forest. Data were collected from the OpenML database, containing medical data of 462 patients and 10 features.

After pre-processing the data using a LabelEncoder for categorical data and a MinMaxScaler for normalization of numerical data, I divided the data into training and test sets with a ratio of 80/20. I then trained the three models on the training data and evaluated their performance on the test data.

The results show that the SVM has the best performance among the three models, with a precision of 78.2% and an accuracy of 77.42%. The ANN model had a precision of 78.9% and an accuracy of 75.41%, while the Random Forest model had a precision of 59.7% and an accuracy of 68.82%. These results suggest that the SVM is the best model for predicting heart disease from these data.

In addition, I evaluated the performance of each model using accuracy as an additional evaluation metric. Accuracy measures the proportion of true positives among all positive predictions. The results show that the SVM has an accuracy of 78.2%, while the ANN model has an accuracy of 78.9%. The Random Forest model has the lowest accuracy of 59.7%. These results show that SVM and ANN are the most accurate models for predicting heart disease.

In conclusion, I successfully predicted heart disease using three Machine Learning algorithms, SVM, ANN and Random Forest. The results showed that SVM and ANN are the most accurate models for predicting heart disease from this data. These results can be used to help medical professionals diagnose heart disease more quickly and efficiently.

B. Summary of results and contributions

The objective of this project was to develop Machine Learning models to predict heart disease. I used three classification algorithms to achieve this goal: SVM, ANN and RandomForest. The results were evaluated using metrics such as precision and accuracy.

The results obtained show that the SVM model has the highest accuracy with 78.2%, while the ANN has the highest accuracy with 75.41%. The RandomForest model gave comparatively lower results with 59.7% accuracy and 68.82% accuracy. Thus, the results indicate that the SVM model is the best for predicting heart disease in this case.

The data were preprocessed before being used to train the models. I used the label encoder and MinMax scaling to normalize the features. The data was then split into training and test sets using the train_test_split function.

The SVM model uses a linear kernel to separate the data into two classes. The ANN model uses a two hidden layer neural network architecture with a ReLU activation function for the hidden layer and a sigmoid activation function for the output layer. The RandomForest model uses a set of several decision trees to make the final decision.

In conclusion, the results of this project showed that Machine Learning models can be successfully used to predict heart disease. The results also showed that the SVM model is the most appropriate for this type of prediction. However, it should be noted that other types of models can be successfully used to solve this problem. This project can be extended by using other classification algorithms and exploring other data preprocessing techniques to improve prediction performance.

C. Future perspectives

The results obtained from the analysis of the three classification models (SVM, ANN and Random Forest) showed that the SVM gave the best result with 78.2% accuracy and 77.42% accuracy. The ANN model gave 78.9% accuracy and 75.41% accuracy, while the Random Forest model gave 59.7% accuracy and 68.82% accuracy. The results show that classification models can be used to predict heart disease with acceptable accuracy.

However, there is still much room for improvement for this type of prediction. First, a larger and more complete database could be used to train the models, which could increase their accuracy. In addition, it would be interesting to add new features such as diastolic blood pressure, HDL cholesterol level, triglyceride level, blood glucose level, and body mass index (BMI) to further improve the accuracy of the prediction.

On the other hand, it would be possible to use more advanced Machine Learning techniques such as deep neural networks or reinforcement learning algorithms to achieve even better results. In addition, it would be possible to use more advanced data preprocessing techniques such as dimension reduction, feature selection, advanced normalization, etc., to improve the quality of the input data and, consequently, the quality of predictions.

Ultimately, the future prospects for predicting heart disease using Machine Learning models are very promising. There are still many opportunities to improve these models, and their use could have a significant impact on public health by enabling earlier detection and faster management of heart disease.

VII. References

[1] Dey, D., et al. "Integrated prediction of coronary heart disease: assessing its performance in a clinical environment-The Penn State Heart Score." Journal of Cardiovascular Computed Tomography, vol. 2, no. 2, 2008, pp. 89-96.

[2] Choi, E., et al. "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks." Journal of Machine Learning Research, vol. 18, no. 1, 2017, pp. 5442-5468.

[3] Wang, Y., et al. "A Deep Learning Algorithm for Prediction of Coronary Artery Disease from ECG." IEEE Transactions on Biomedical Engineering, vol. 65, no. 11, 2018, pp. 2485-2492.

[4] Vaithiyanathan, V., et al. "A Machine Learning-Based Decision Support System for Coronary Heart Disease Prediction." Proceedings of the 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), IEEE, 2018, pp. 211-216.. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955 *(references)*

[5] Alaa, A.M., van der Schaar, M. "Prognostication and Risk Factors for CVD using Machine Learning". JACC. Cardiovascular Imaging. 2020; 13(11): 2336-2357.

[6] Dey, D., Dey, M., Jiang, Y., et al. "Machine learning predicts individualized cardiovascular risk reduction in myocardial perfusion imaging." JACC Cardiovasc Imaging. 2019; 12(11 Pt 2): 2227-2229.

[7] Choi, E., Bahadori, M.T., Sun, J., et al. "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks". Journal of Machine Learning Research. 2016; 56:301-318.

[8] Wang, S., Zhou, J., He, Y., et al. "A multi-label classification approach for the prediction of coronary artery disease based on ECG signals." BMC Medical Informatics and Decision Making. 2019; 19(1):248.

[9] Vaithiyanathan, V., Hussain, A., El-Emam, A., et al. "A machine learning approach for predicting coronary artery disease from cardiac computed tomography angiography images". International Journal of Computer Assisted Radiology and Surgery. 2018; 13(3):389-399.

[10] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., & Vlahavas, I. (2018). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 16, 97-105.

[11] Wang, Z., Liu, J., & Zuo, W. (2018). A machine learning-based framework for heart disease diagnosis. Journal of medical systems, 42(8), 139.

[12] Acharya, U. R., Fujita, H., Lih, O. S., Adam, M., Tan, J. H., Chua, C. K., & Lim, C. M. (2017). Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. Information Sciences, 405, 81-90.

[13] Zreik, M., Lessmann, N., van Hamersvelt, R. W., Voskuil, M., Wolterink, J. M., Išgum, I., & Leiner, T. (2018). A deep learning framework for segmentation of cardiac left ventricle anatomy. Scientific reports, 8(1), 1-10.

[14] Bai, W., Suzuki, H., Qin, C., Tarroni, G., Oktay, O., Matthews, P. M., & Rueckert, D. (2018). Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of cardiovascular magnetic resonance, 20(1), 1-13.

[15] Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2018). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657-2664.

[16] Dey, D., Gaur, S., Ovrehus, K. A., Slomka, P. J., Betancur, J., Goeller, M., & Berman, D. S. (2018). Integrated prediction of lesion-specific ischaemia from quantitative coronary CT angiography using machine learning: a multicentre study. European Radiology, 28(6), 2655-2664.

[17] Ozturk, S., & Sengur, A. (2018). Computer-aided diagnosis of arrhythmia using PCA, LDA, ICA and discrete wavelet transform. Measurement, 117, 107-115.

[18] Zhu, H., Wang, S., Zhou, S., & Huang, Y. (2019). A Convolutional Neural Network Based Method for Automatic Detection of Coronary Artery Disease in CT Angiography Images. Journal of healthcare engineering, 2019.