The Pennsylvania State University

The Graduate School

College of Engineering

# APPLICATIONS OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING IN HEALTHCARE: A CARDIOVASCULAR CASE

A Thesis in

Industrial Engineering

by

Hassan Alamoudi



Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2019

The thesis of Hassan Alamoudi was reviewed and approved* by the following:

Soundar Kumara
Allen E. Pearce/Allen M. Pearce Professor of Industrial Engineering
Thesis Advisor

Hui Yang
Associate Professor of Industrial and Manufacturing Engineering

Robert Voigt
Industrial and Manufacturing Engineering
Professor and Graduate Program Coordinator

*Signatures are on file in the Graduate School

**ABSTRACT**

Artificial Intelligence (AI) is becoming a ubiquitous term that is used in many fields of research or the popular culture. Among these fields that was affected by this hype is the healthcare sector. Along with its subdomain, Machine Learning (ML), they established an environment of interest in the promises of machines versus humans' capabilities. Though artificial intelligence applications in healthcare such as interpreting ECGs could date back to the mid of the twentieth century, the promises of AI still at its beginning when it comes to new breakthroughs. This is due to the transformation into a digital world and new advancements in the processing capabilities. Computer vision has contributed the most to the healthcare sector where it can leverage doctors and practitioners with automated classification and annotations as a preparing step. This kind of mechanism is the best suited for applications of AI in healthcare. However, the amount of data in other forms such as textual or lab results is exceeding the force power. While a solution could be to use machines to learn and propose solutions, the results could be catastrophic and human lives are on stake. So, explainable AI could be beneficial where it analyzes and makes predictions that can be trusted by the users. The study here is conducted on cardiovascular patients' dataset to predict the presence or absence of the disease. Classifications techniques used include Naïve Bayes, Logistic Regression, Decision Trees, Support Vector Machines, and Artificial Neural Networks. The Logistic regression model achieved the best Area under the curve. Moreover, an extension of the previous studies discussed is conducted to explain the model and to show how models of AI can be trusted and not used as black-boxes.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

## ACKNOWLEDGEMENTS

First and Foremost, I thank Allah for his grace over me. Secondly, I would like to thank my advisor Dr. Soundar Kumara, Professor of Industrial Engineering, for his valuable support and exemplary guidance throughout the course of my thesis. I also cannot thank him enough for the sheer knowledge he provided me through the teachings of his courses.

I would also like to express my gratitude to Dr. Hui Yang, Associate Professor of Industrial and Manufacturing Engineering, for taking his time to review my thesis and giving constructive feedback.

Also, I express my profound gratitude to my family and friends for their continuous support and encouragement throughout my graduate studies. This would not have been possible without all of you.

**Chapter 1**

# Introduction

Artificial intelligence is a buzzword nowadays because of its involvement in many of the recent advances in science and technology. Different applications of AI and its subtype, machine learning, have seen the light because of the tremendous amount of available data and improved processing power. The capabilities of AI provide a potential solution for the necessities of the healthcare sector where an abundant amount of data is generated through examination notes, lab tests, and medical devices that is faced with a shortage of human personnel who can analyze them and provide the medical care needed. Thus, as the different methods of AI process and analyze the available data, they can be used to assist different clinical practices. The literature of AI applications in healthcare can be further discussed through the type of data in addition to the special field of practice in medicine.

Cardiovascular diseases are accounted for an estimated 31% of all deaths worldwide according to World Health Organization (WHO). One of the major causes of death requires an attention if new technologies are developed and can be of help. (Jiang et al. 2017) support this where they found out that the amount of published papers of AI in cardiovascular literature on PubMed in the duration between 2013-2017 reached 1000 papers and came as the third type of diseases that is discussed in AI literature in healthcare.

Though applications of AI in healthcare are evolving and are expected to augment experts' decisions, they suffer sometimes from what is known as a black-box problem. The model takes an input and provides an output without the user involvement in the prediction analysis. This will lead to a trust problem and sometimes catastrophic results when the application is of vital importance such as medical applications and replacing humans with machines.

## Motivation

Cardiovascular diseases (CVDs) are defined by World Health Organization as the disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. It is estimated that 18 million deaths occur worldwide due to CVDs. Therefore, it is essential to diagnose those with the signs of CVDs and intervene early so their health will not deteriorate.

AI systems can be trained on available data to learn from the patterns within them. Learning can be tailored using different algorithms and various configuration of parameters. Moreover, metrics that examine the models' performance such as accuracy can be monitored and reported.

Previous Studies conducted are usually focused on the accuracy of models in the early prediction of the disease. Though the early detection of heart problems could be beneficial, decisions should not be based merely on black-box models or performance metrics only. Experts and patients who represent the end users of the AI systems should have trust in these systems. Trust can be gained by the sound predictions of the model and being in alignment with expert practices in supervised environments.

In light of all the above factors, it is beneficial to have a classification system that can classify cardiovascular patients into healthy or unhealthy based on simple medical signs. If the predictions of the system are trustworthy by the users, it gives them a reason to depend on the system and believe in its outcomes.

## Problem Definition

The goal of this paper is to develop a classification system that can accurately predict and justify presence or absence of a heart problem in patients. The problem is a binary classification problem that provides medical signs of patients and can be approached using AI classification algorithms. The binary classes will be defined as - ‘1’ being ‘Healthy and ‘0’ being ‘Unhealthy’. Features available of the data include demographical and medical vital signs. The algorithms that will be used for classification purpose are Multinomial naïve Bayes, Logistic regression, Decision Trees, Support Vector Machine, and Artificial Neural Networks. After that, the best model will undergo an explainable phase where different measures will be examined.

## Organization of the Thesis

The rest of this thesis is structured as follows: chapter 2 provides a literature review of the applications of AI and ML in healthcare in general and cardiovascular in particular as long as a brief explanation of the used algorithms. Chapter 3 discuss the dataset in hand. Methodology is discussed in chapter 4. Chapter 5 provides the results and the analysis accompanying them. Lastly, chapter 6 contains the conclusion of this paper and discuss the potential of future work.

# Chapter 2

# Literature Survey and Background

## Applications of AI and ML in Healthcare

AI applications in healthcare could be divided based on the clinical activity and also based on the type of data. Some of the forms that the data can come in include demographics, medical notes, medical scans, pathology slides, skin lesions, retinal images, electrocardiograms, vital signs, electronic recordings, physical examinations and clinical laboratory and images. AI methods are trained using these different types of data and usually are compared with physicians' assessment based on the area under the curve (AUC) obtained from the plot of true-positives versus false-positive rates known as the Receiver Operating Characteristic (ROC) to validate their suitability. Eric Topol (2019)

This section aims to provide the reader of some of the applications of AI in healthcare and is not intended to be an extensive overview. Therefore, the following sections are divided into different applications of AI based on the type of data.

### Image Analysis

Biomedical images are considered one of the forms of medical data and they represent the most studied form of data in AI healthcare literature. Part of this is related to abundance of medical imaging scans that are stored in many modern hospitals' Picture

Archiving and Communication Systems (PACS). The other part is related to the advancement of computer vision algorithms, especially convolutional neural networks.

In radiology, for example, chest X-rays represents the most common type of medical scans with more than 2 billion performed per year worldwide. Chest X-rays can be used in the diagnosing stage to detect many lung-related diseases. Wang et al (2017) applied a 121-layer convolutional neural network to detect pneumonia in 112,000 frontal chest X-ray images. They concluded that their network performed better than four radiologists. However, this result is not optimal when the achieved accuracy is 76% and radiologists usually scan the images for more than detecting pneumonia only. Another study conducted by Li et al (2018) utilized the same dataset mentioned earlier to classify the images into 14 different thoracic diseases. Their accuracy scores ranged from 67% in pneumonia to 91% in emphysema.

In pathology, whole slide imaging (WSI) works as the digitized version of the glass slides. Ehteshami et al (2017) applied different deep learning algorithms to whole slide images to test the accuracy of the algorithms in detecting metastases in tissue sections of lymph nodes of women with breast cancer and compared it against 11 pathologists' diagnoses. The test set consisted of 129 whole slide images and the pathologists had less than one minute for review per slide. Therefore, the results varied in favor of some of the algorithms. While the AUC for the algorithms ranged from 0.556 to 0.994, the mean AUC of the pathologists was 0.810 with the time constraint. On the other hand, the top five algorithms had a mean AUC similar to the pathologist interpreting the slides in the absence of time constraints. The top five algorithms achieved a mean AUC of 0.960 against 0.966 for the pathologist without the time constraint.

Other medicine branches such as dermatology and ophthalmology utilized deep neural networks to detect and classify different diseases related to the different body systems. Esteva et al. (2017) fine-tuned a Google Inception CNN architecture on around 130,000 images of skin cancer depicting 2032 different diseases and concluded that their CNN matched the results of 21 board-certified dermatologists in classifying different kinds of skin cancer.

In ophthalmology, emphasizing on the fact that diabetic retinopathy impacts more than 90 million individuals worldwide and is a major cause of adult blindness, AI offered a real solution to overcome this problem. This solution is claimed to be as the first FDA approved autonomous AI diagnostic system to detect more-than-mild diabetic retinopathy in adults who have diabetes mellitus. The system introduced by Abràmoff et al. (2018), known as IDx-DR, undergo a trial conducted on 900 patients at primary care clinics and kept on trial on an autodidactic mode and then was locked for testing. During the trial, the system surpassed FDA regulations and achieved sensitivity, true positive rate, of 87% and specificity, true negative rate, of 91% for the 819 patients with analyzable images. However, the regulation from the FDA constitutes the system to stop its learning function and be treated as a non-AI diagnostic system which is hindering to its potential.

**Biomedical Electrical Signal Analysis**

Another valuable source of information in healthcare is the data represented as physiological signals coming from the different sensors put on the body skin. Electromyogram (EMG), electroencephalogram (EEG), electrooculogram (EOG) and

Electrocardiogram (ECG) produce different kinds of physiological signals that can be used as inputs for the applications of AI systems in disease detection and diagnosis. Bote-Curiel et al. (2019)

Electromyogram (EMG) is used to evaluate and record the electrical activity produced by skeletal muscles such as muscle state, activation of the muscle and the force generated. These different parts are overlapped in the EMG signal which leads to the problem of classifying the signal into the different parts. Hence, applications of deep learning are being utilized in this field. Faust et al. (2018). For example, limb movement estimation. Xia et al (2017), gesture recognition, Geng et al. (2016), and hand movement classification, Atzori et al. (2016)

Electroencephalography (EEG) is an electrophysiological monitoring method to record electrical activity of the brain. It works by summing up the charges coming from the neurons in the brain that would be leading to an action. The generated signal has a noise characteristic that makes it hard to be interpreted. Furthermore, the brain-computer interface used to record the signals is more suitable for an automatic decision making system. Hence, AI systems has more potential than human practitioners. Applications found in this field include sleep-state identification, Fraiwan and Lweesy (2017), seizure detection, Acharya et al. (2018), and emotion classification, Zheng et al. (2014).

Electrooculography (EOG) is a technique for measuring the corneo-retinal standing potential that exists between the front and the back of the human eye. It is used to detect the movement of the eye and thus is helpful in ophthalmological diagnosis. However, the signals are affected by noise and thus become hard to interpret. Different AI applications are used to deal with these obstacles with previous works found in drowsiness detection, Zhu et al. (2014), driving fatigue detection, Du et al. (2017), and Sleep stage classification, Xia et al. (2015).

Electrocardiogram (ECG) is a recording of the electrical activity of the heart using electrodes placed on the chest. The signals of the ECG can be divided to different intervals. These intervals are interpreted to diagnose the activity of the heart. During the interpretation, the practitioner looks for morphological changes that would indicate a specific cardiac problem. These morphological changes could be short in duration as transients or could be present all the time. AI systems have been utilized for the automatic detection of different cardiac abnormalities. Previous work includes coronary artery disease detection, Acharya et al. (2018), irregular heartbeat classification, Majumdar and Wardand (2017), congestive heart failure detection, Zheng et al (2014).

**Electronic Health Records Analysis**

Electronic Health records (EHR) are considered the primary carrier of health information. They contain essential information that can be structured (e.g. diagnosis, medications, laboratory tests) and unstructured (e.g. free-text clinical notes) data. The abundant and growing amount of medical data gives the opportunity to AI systems to become global players in healthcare and medicine by developing new applications to classify diseases and their subtypes accurately, to create personal patient treatments or to guide the development of new therapies. Most of the work done recently is concentrated on applying deep learning techniques in a supervised environment to conduct a predictive study. However, Unsupervised models can be used to conduct a descriptive study to discover unknown patterns between EHR's.

The challenges that are facing utilizing different AI systems include the nature of the data itself. EHR's are characterized as high-dimensional, temporal, sparse, irregular and bias. To overcome these challenges, Cheng et al. (2015) proposes a deep learning approach for feature extraction, electronic phenotyping, from patient EHRs. They used a four-layer convolutional neural network model for extracting phenotypes and perform prediction on congestive heart failure and chronic obstructive pulmonary disease and showed significant advantages over the baselines which was a logistic regression model. The results of this study showed a slight improvement of the CNN models over the logistic regression. However, this has happened on the expense of interpretation.

Another application to exploit the EHR records more efficiently is to use them for the purpose of personalized treatment care. This is the purpose of DeepCare, an end-to-end dynamic neural network utilizing recurrent neural network (RNN) with long short-term memory (LSTM) that reads medical records, stores previous illness history, infers current illness states and predicts future medical outcomes. The study has been conducted on diabetes and mental health and evaluated on disease progression modeling, intervention recommendation and future risk prediction, Pham et al (2017).

Moreover, application of natural language processing (NLP) can be utilized here either to extract useful information or to study the similarity between medical concepts. For example, Afzal et al. (2017), developed an NLP system for automated discovering of peripheral arterial disease (PAD) cases from clinical narrative notes and compared the performance of the NLP algorithm with billing code algorithms. By exploring a variation of natural language processing models that can learn on concepts taken from structured ontologies and extracted from free-text, a semantic similarity between medical concepts represented by journal abstracts and patient records is measured in the study by Vine et al. (2014) They concluded that their results correlate with expert human assessors and perform better than some of state-of-the-art benchmarks for medical semantic similarity. Table 2-1 summarizes the studies reviewed in this section divided based on the type of data and the AI application conducted.

Table **2-1**: Summary of the study reviewed in this section divided based on the type of data and the application carried

| Data | | Authors | Study | Application |
|---|---|---|---|---|
| Medical Imaging | | Wang et al. (2017) | detect pneumonia in chest X-ray images | CNN |
| | | Li et al. (2018) | classify chest X-rays into 14 different thoracic diseases | CNN |
| | | Ehteshami et al. (2017) | detecting breast cancer | SVM, RF, and CNN |
| | | Esteva et al. (2017) | Detecting skin cancer | CNN |
| | | Abràmoff et al. (2018) | Diabetic retinopathy diagnosis | CNN |
| | | | | |
| Medical Signals | EMG | Xia et al. (2017) | Limb movement estimation | RNN |
| | | Geng et al. (2016) | Gesture recognition | CNN |
| | | Atzori et al. (2016) | hand movement classification | CNN |
| | EEG | Fraiwan et al. (2017) | Sleep-state identification in newborns | Autoencoders and CNN |
| | | Acharya et al. (2017) | Seizure detection | CNN |
| | | Zheng et al. (2014) | Emotion classification | DBN |
| | ECG | Acharya et al. (2017) | Coronary artery disease detection | CNN |
| | | Majumdar et al. (2017) | irregular heartbeat classification | Robust deep dictionary learning |
| | | Zheng et al. (2014) | congestive heart failure detection | CNN |
| | EOG | Zhu et al. (2014) | Drowsiness detection | CNN |

| | | | | |
|---|---|---|---|---|
| | EOG and ECG | Du et al. (2017) | Driving fatigue detection | Autoencoder |
| | EOG, EEG | Xia et al. (2017) | Sleep stage classification | DBN |
| | | | | |
| EHR | | Liu et al. (2015) | Prediction of congestive heart failure and chronic obstructive pulmonary disease | CNN |
| | | Pham et al. (2016) | DeepCare: a dynamic neural network with memory for prediction based on patient history | RNN with LSTM |
| | | Afzal et al. (2017) | Automated discovery of peripheral arterial disease (PAD) cases based on clinical notes | NLP |
| | | Vine et al. (2014) | Medical Semantic Similarity with a Neural Language Model | NLP |

The previous studies reflected upon how different AI applications are implemented in healthcare. Different data types such medical image, electrical signals, and medical records were among the different types that AI systems exploited to create a pathway in healthcare. Many various applications could be found in the literature and encompassing all of them in one paper is not feasible. Therefore, other published papers focused on reviewing applications based on diseases related to different body systems. Accordingly, this paper focuses on cardiovascular diseases and thus the next section will attempt to provide a review on the applications done in cardiovascular medicine with a focus on one dataset to preserve consistency.

**Applications of AI and ML in Cardiovascular Medicine**

As stated before, AI systems have the potential to achieve early diagnosis by exploiting the rich data found in the different forms such as medical images, electrical signals, and electronic health records. Medical images are suitable for applications of computer vision, but they would not be the primary tool to diagnose a CVD. Notably, A study conducted on diagnosis of chronic myocardial infarction using MRI scans can be found in the work of Zhang et al. (2019). Moreover, biomedical electrical signals are discussed earlier in the section of application on ECG. Interestingly, diagnosis of CVDs usually could be demonstrated by raised blood pressure, glucose, and lipids and other signs such as overweight and obesity that could be found in electronic health records. Thus, researchers have been developing different prediction systems utilizing these records.

One of the famous datasets is Cleveland dataset. The dataset that is concerned with this paper. A further discussion will be given in the next section about it. Researches have been utilizing the Cleveland dataset to test the accuracy of their prediction models. Other specialists can use it along with datasets they possess from their hospitals to conduct their studies. Their results and approaches differ accordingly. Therefore, this section will try to provide a review about the studies conducted using this dataset.

Medhekar et al. (2013) presented a Naive Bayes classifier for the detection of heart disease and showed how it can be used for classification purposes. They classified the medical data into five categories related to the heart disease risk prediction. Namely, no, low, average, high, and very high risk. They trained their system on the Cleveland dataset with different number of instances for training. It is not obvious how the holding-out was done and therefore this study may suffer from overfitting. The best reported accuracy was 89.58% with 25 misclassified instances without any further explanation of the misclassification.

Another implementation of Naïve Bayes classification was conducted by Vembandasamy et al. (2015). The authors suggested a heart diseases prediction system (HDPS) based on the datamining approaches. Their work was carried in Java using the WEKA environment. The researchers implemented their study on 500 patients' clinical data collected from a diabetic research institute in Chennai, India. The dataset contained a feature resembling the possibility of heart disease. The reported results of the Naïve Bayes classifier show that it was accurate on 74% of the instances.

Gudadhe et al. (2010) performed a classification study using SVM and MLP and obtained 80.41% accuracy on the former and 97.5% in the latter. It is worth mentioning that the neural network was trained on the five classes of the heart disease rather than absence or presence of the disease as it is the case in the SVM training.

Patel et al. (2015) compared different algorithms of Decision Tree classification to find the best performance in CVDs diagnosis. They used algorithms like J48 algorithm, Logistic model tree (LMT) algorithm and Random Forest (RF) algorithm using WEKA. The training and testing were conducted using 10-fold cross validation. Their models contained five classes to classify the likelihood of heart disease. The results of the study were not optimistic since the best algorithm, J48 algorithm, outperformed the rest with an accuracy of 56.76%.

Furthermore, Sabarinathan and Sugumaran (2014) used parameters such as age, gender, chest pain, heart rate achieved to classify heart disease using decision trees with J48 algorithm. Their dataset contained 240 instance that was split in half for training and testing. They achieved an accuracy of 75.83% using all the features. After that, the accuracy is improved to 85% when more irrelevant features were removed. They concluded that thalassemia, chest pain type and number of major vessels are the primary attributes that would help the classification. However, there was no reporting of this conclusion and how does it match with the experts.

Kahramanli and Allahverdi (2008) applied a hybrid neural network between Fuzzy Neural Network and Artificial Neural Network and obtained an accuracy of 86.8% on the Cleveland dataset. The fuzziness aspect is implemented here since it is based on the fuzzy set and fuzzy logic. These concepts add to the traditional clustering where they allow patterns to belong to more than one cluster with different degrees.

Based on the same principle of neural networks ensemble, Das et al. (2009) developed a decision support system with SAS base software to help physicians in diagnosing heart diseases. Their neural network ensemble consisted of three models and obtained a classification accuracy of 89%. However, this work was based on the Cleveland dataset and did not address how physicians could incorporate their own datasets of patients.

The study of Zhou and Jiang (2004) proposed a newer approach to tackle the problem of achieving a middle ground between retaining comprehensibility of decision trees and generalization ability of neural networks. Their proposed neural network ensemble based on C4.5 algorithm, NeC4.5, would employ neural networks at the beginning to generate newer training examples that can be fed to the C4.5 for classification. On the Cleveland heart dataset, the authors showed that their newer approach was able to improve prediction error significantly when they added new training examples though it also resulted in a larger tree with more nodes. On the other hand, their result was not significant when the ratio of new to existing examples parameter was set to zero. Although this approach could be promising, the training of the neural network ensemble incurs a time cost that cannot be ignored.

## Accuracy Versus Interpretability

It can be seen from the previous works that as more deep learning is integrated, the accuracy of the model increases, and this is happening on the expense of interpretability. However, in fields like medicine, the mere suggestion of using models as black boxes is not acceptable. Practitioners and patients have the right to know why the model behaved the way it did and how could it justify its decisions. This will establish trust and support actions taken on such predictions. Therefore, Ribeiro et al. (2016) proposed an explanation technique called Local Interpretable Model-Agnostic Explanations, or LIME, that works on two levels of trust to explain machine learning or even deep learning models.

LIME approach introduces solutions to explain single predictions and the whole model. Single predictions trust means the user can trust that prediction and is willing to act based on it. This is quite applicable in highly sensitive environments such as medicine or counterterrorism where misclassifications will result in catastrophic outcomes. On the other hand, trusting the whole model means the belief that the model will behave in a reasonable way if deployed. It is noted this is reflected in evaluation metrics such as accuracy, but real-world data is often significantly different. Thus, inspection of different individual predictions is a worthwhile solution.

By "explaining a prediction", the authors refer to the use of textual or visual artifacts that provide qualitative understanding of the relationship between the instance's components (e.g.  words in text, patches in an image) and the model's prediction. Figure 2-1 explain the process clearly where the explanation provides the user, a doctor in this case, with relative weights of the variables, symptoms here, it used to build its decision. The green color is for symptoms that are present and supporting the prediction and the red ones are for evidence against it. The doctors with their previous knowledge can decide on whether or not to trust this prediction. This process is way more appropriate than presenting a prediction without any evidence. Furthermore, if there were different explanations of different instances, the model itself can be trusted to act reasonably.

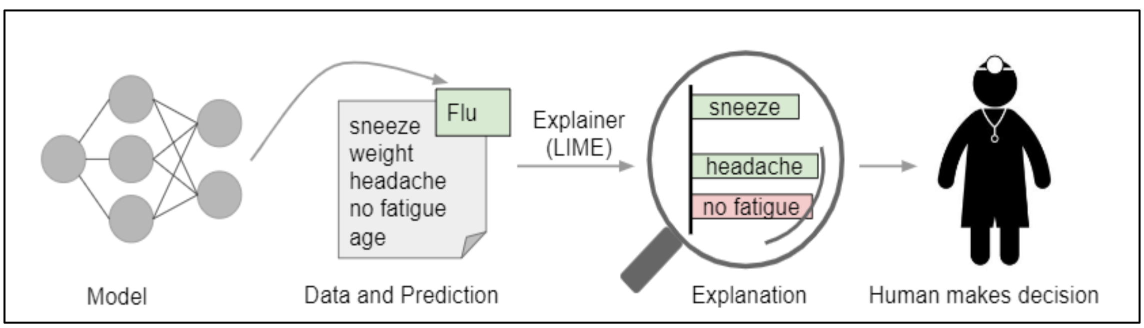


Figure **2-1**:  Explanation of a prediction using a model to predict a medical condition

## Machine Learning Algorithms for Classification

### Naïve Bayes

The Naive Bayes classifier is based on the Bayes' theorem of conditional probabilities. It works by calculating the probability for a class depending on the value of the feature over all the features. The naïve term comes from the assumption of independence of the features. The conditional probability of a class $C_k$ is calculated as:

$$P(C_k|x) = \frac{1}{Z} P(C_k) \prod_{i=1}^{n} P(x_i|C_k)$$

Z is a scaling parameter that ensures the sum of probabilities for all classes is 1. The conditional probability of a class is the class probability times the probability of each feature given the class, normalized by Z.

Naive Bayes is an interpretable model because of the independence assumption. For each feature, it is very clear how much it contributes towards a certain class prediction, thus interpreting the conditional probability.

### Logistic Regression

Logistic regression models the probabilities of the target belonging to a certain category. The basic model works on binary classification problems. It is an extension of the linear regression model for classification problems. The logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. The logistic function is defined as:

$$\log(\eta) = \frac{1}{1 + \exp{(-\eta)}}$$

Interpretation of weights in logistic regression differs from linear regression because the outcome in logistic regression is a probability between 0 and 1. A change in a feature by one unit increases the log odds ratio by the value of the corresponding weight.

**Decision Trees**

When the relationship between features and outcome is nonlinear or where features interact with each other, linear regression and logistic regression models' performance deteriorate. A simple solution to these situations is decision trees. Decision trees works by splitting the data multiple times based on a measure such as information gain that determines how much information can be gained by that split. The splitting procedure results in different subsets of the dataset. The final subsets are called leaf nodes. The prediction is occurring on these leaf nodes where the predicted outcome is the average outcome of the training data in that subset.

There are different algorithms for growing a tree and they differ in how the tree is structured. A mathematical description of how the relation between an outcome y and features x is reflected as:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} C_m I\{x \in \ R_m\}$$

$R_m$ is the leaf node subset and $I_{\{x \in R_m\}}$ is the identity function that returns 1 if the instance belongs to that subset or 0 otherwise. if an instance falls into a leaf node $R_l$, the predicted outcome will be $\hat{y} = c_l$, where $c_l$ is the average of all training instance in $R_l$.

Interpretation of decision trees is quite forward. Following the edges of the tree from the root node to the leaf node will explain how a certain prediction is made. However, this fact holds true as long as the trees are short. The deeper the tree, the harder it gets to understand the decision rules of the tree.

**Support Vector Machine (SVM)**

Given labeled training data, SVM works by finding a hyperplane that separates classes with the biggest gap (maximum margin) on either side using an optimization technique that can be solved with a quadratic programming. In a two dimensional space, this hyperplane would be a line the separate the plane to two parts. However, when the training data are cluttered, it becomes hard to separate and classify them by a single line. Therefore, SVM transforms the data into a higher feature space using a kernel trick and afterward classify the data.

**Artificial Neural Networks**

Attempting to imitate human brains, ANNs have continuously discovered broad applications in answering an expansive scope of nonlinear problems and have drawn expanding considerations of research community. The most recognized feature of ANNs is simply the nonlinear, nonparametric, data-driven, and adaptable nature. ANNs do not require knowing information about the underlying statistical distributions of the data. They adaptively build the fitted model from the available information within the data, learn from training examples, and afterward generalize the acquired knowledge to predict the nature of future events.

ANNs models' performance is affected by different number of parameters. These parameters include but not limited to proper selection of network architecture, training algorithm, number of hidden layers, number of nodes in each layer, and activation functions. Of these different tasks, selection of network training algorithm perhaps could be considered as the most important task in ANN modeling. Among the different algorithms suggested so far, backpropagation that was developed by [Rumelhart et al., 1986] could be considered as the best training method. It works by updating the network weights in the direction of the decrease of the error function, which is known also as the gradient steepest descent method.

The interpretation of both ANN and SVM – in nonlinear situations – can be considered as dealing with a black box. For example, a neural network would perform thousands of calculations based on the chain rule to update a weight in a layer in the network. Following the exact mapping for these calculations is quite unfeasible for us humans.

## Chapter 3

# Overview of the Data

### Dataset Description

The dataset used in this study is the Cleveland heart disease dataset collected by Robert Detrano (1988), found in the UCI machine learning repository. It consists of 13 variables measured on 303 individuals. The 14$^{th}$ variable, called target, is a binary variable that signals the presence of heart disease or not. The variables and their descriptions are discussed in Table 3-1 below

Table **3-1**: Cleveland Dataset Features Description

| Variable | Description | Type |
|---|---|---|
| Age | Age in years | Integer |
| Sex | 1=Male, 0=Female | Binary |
| Cp | cp: chest pain type<br>0: asymptomatic<br>1: atypical angina<br>2: non-anginal pain<br>3: typical angina | Categorical |
| Trestbps | Resting blood pressure (in mm Hg on admission to the hospital) | Continuous |
| Chol | Serum cholesterol in mg/dl | Continuous |
| Fbs | (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) | Binary |
| Restecg | Resting electrocardiographic results<br>0: showing probable or definite left ventricular hypertrophy by Estes' criteria;<br>1: normal;<br>2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) | Categorical |

| | | |
|---|---|---|
| Thalach | Maximum heart rate achieved | Continuous |
| Exang | Exercise induced angina (1 = yes; 0 = no) | Binary |
| Oldpeak | ST depression induced by exercise relative to rest | Continuous |
| Slope | The slope of the peak exercise ST segment<br>0: downsloping; 1: flat; 2: upsloping | Categorical |
| Ca | number of major vessels (0-3) colored by fluoroscopy *(4 missing values) | Integer |
| Thal | Thallium stress test result 1 = fixed defect; 2 = normal; 3 = reversable defect) *(2 missing values) | Categorical |
| Target | Presence of heart disease:<br>0 = disease, 1 = no disease | Binary |

Table 3-1 explain the features of the dataset after it has been processed for production. Six instances out of 303 were dropped due to missing values and one instance was dropped because it was a duplicate one. Classes of categorical data are processed from the original classes to the explained ones.

# Chapter 4

## Methodology

The goals of this paper are as follows:

1. Performing an exploratory analysis of the dataset.
2. Conducting different classification experiments between traditional machine learning methods and advanced deep learning techniques.
3. Comparing between the previous experiments based on interpretability-accuracy criteria by applying Ribeiro’s method.

### Exploratory Analysis

First step in exploring this dataset is to see the number of patients in each class. Figure 4-1 shows that out of the 296 patients 160 belongs to class ‘1’ which means they are healthy while 136 are not and prone to develop a heart disease. This dataset seems balanced in terms of representing both classes.


Figure **4-1**: Distribution of Target classes; 1: Healthy, 0: Not healthy

Figure 4-2 examine the correlation between the variables. The existence of high correlation means that the variables are redundant.

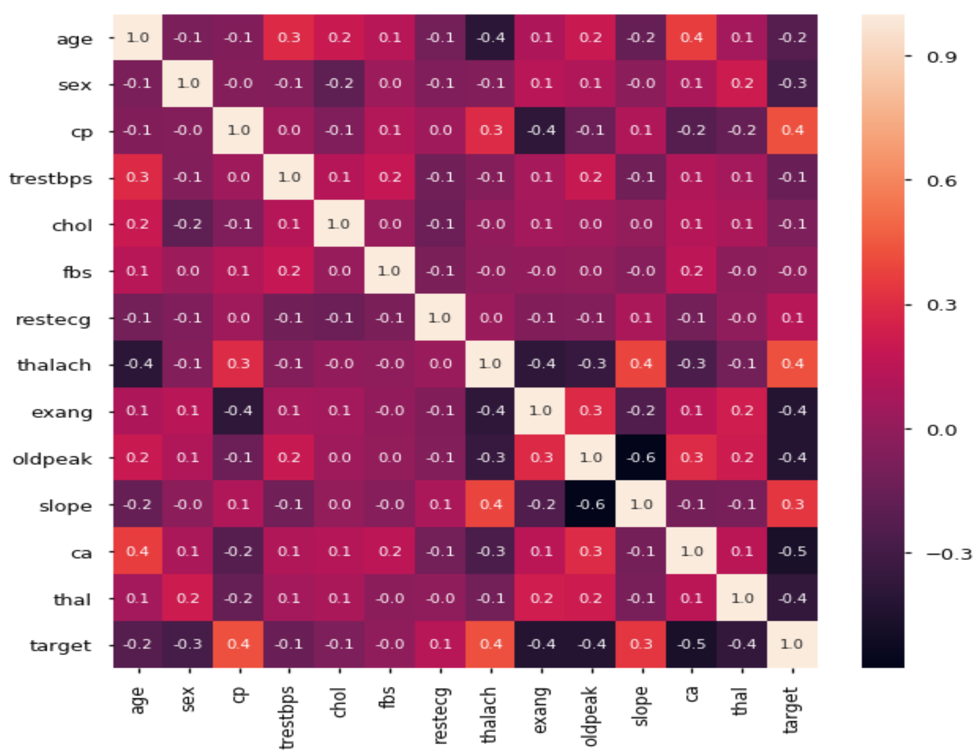


Figure **4-2**: Correlation matrix between the variables

The correlation matrix in figure 4-2 does not show any strong correlation between any of the variables. This means that all the variables could contribute differently to the classification of the disease and thus cannot be ignored. Therefore, next figures will try to explore how both classes of the target relate to the different variables.

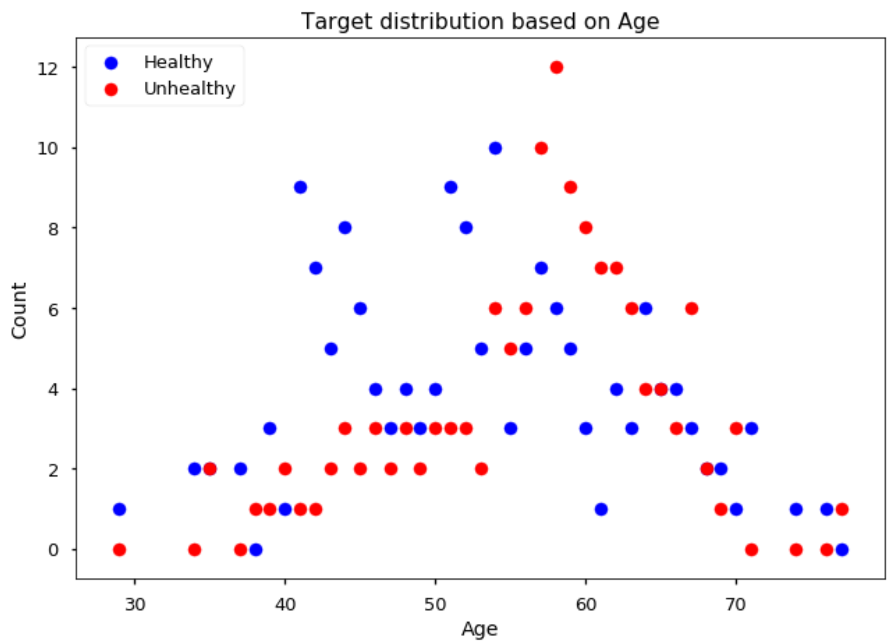


Figure **4-3**: The relation between the variable Age and the classes of the target

Figure 4-3 Shows the relation between the *age* of the patient and the condition of the heart disease. It could be inferred that the likely of the disease increases with age.

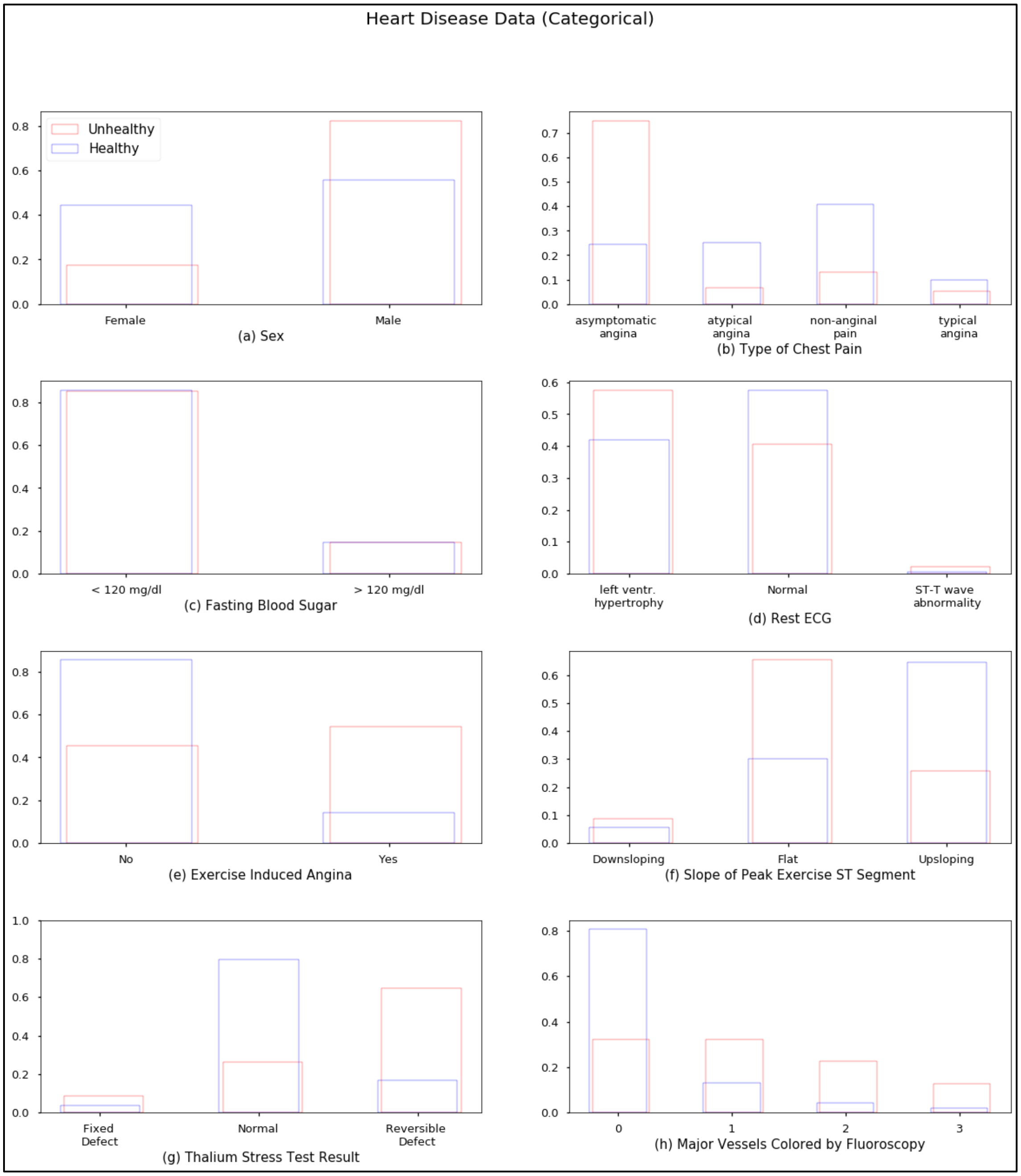


Figure **4-4**: (a) to (h) Relation of different categorical variables with the target

Figure 4-4 shows how the different categorical variables' classes relate to the diseases based on the prevalence of the disease in the dataset, 160 healthy and 136 unhealthy. For example, 80% of the unhealthy are from the male population while 20%

are female. Moreover, most of the unhealthy individuals have asymptomatic anginal chest pain and so on. On the other hand, figure 4-5 shows how both classes of target react with different values of the continuous variables. In conclusion, it can be inferred from these figures that people diagnosed with heart disease are more likely to be older males who have higher blood pressure and higher cholesterol levels and so on than people who do not have a heart problem. Such graphs are useful to get a general idea of the dataset and how each of the features would be useful in a machine learning model. However, the next section will apply different classifiers using all the attributes of the dataset leading up to the final explainable model that will show how a prediction is made.

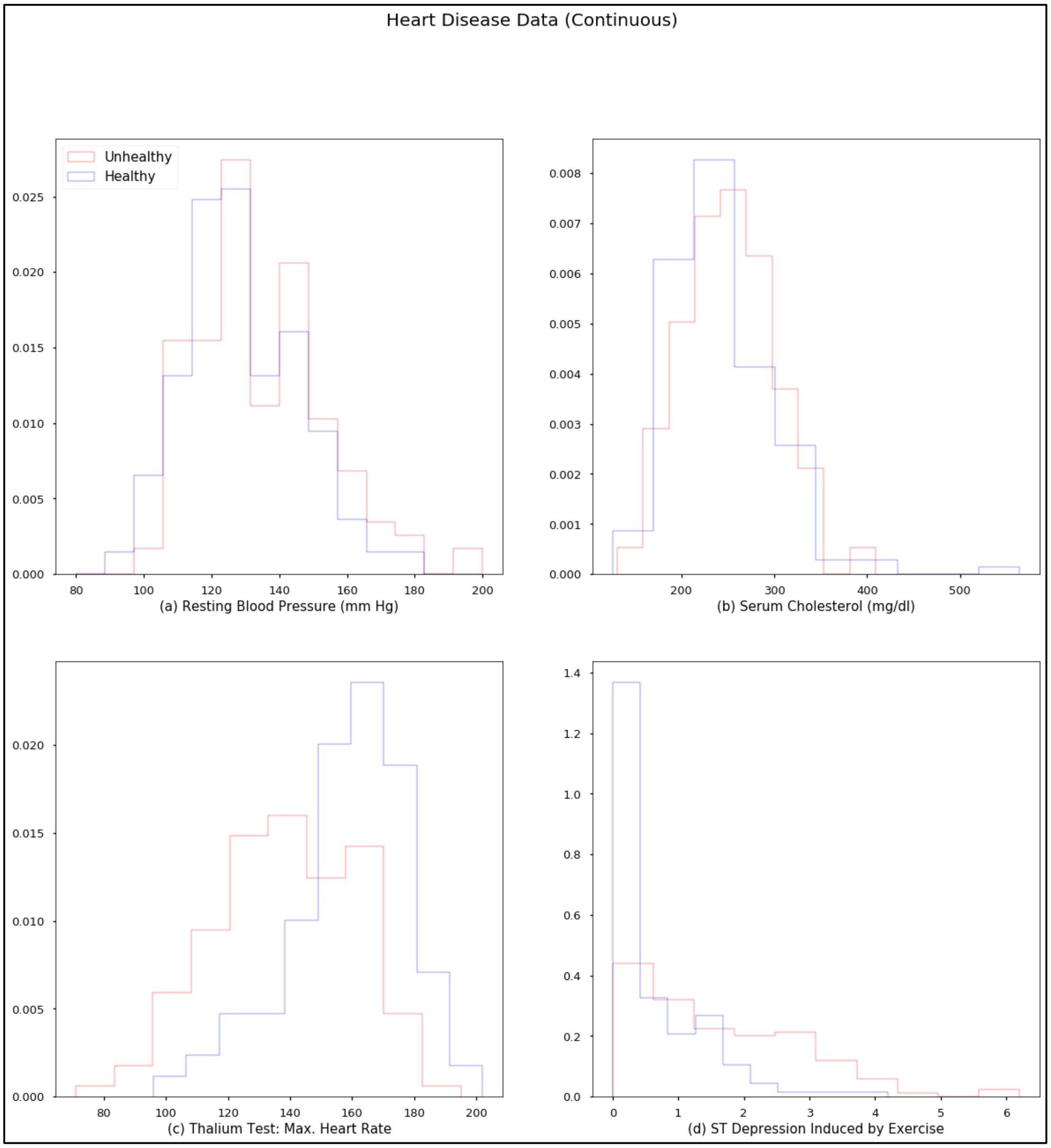


Figure **4-5**: (a) to (d) The relation of the continuous variables with the target

## Chapter 5

# Analysis and Results

This section discusses the analysis of the results obtained using the different classification methods conducted on the prediction of the presence of the heart disease. Based on the results, a following part will discuss the explainable model that is based on the highest achieved accuracy.

## Classification Techniques

### Naïve Bayes

For Naïve Bayes, the selected method was a Mixed Naïve Bayes since it can treat the features each by its own distribution. Thus, categorical features will be treated as coming from their categorical distribution and continuous data as coming from a gaussian distribution after they have been normalized. Figure 6-1 shows the confusion matrix of the MixedNB.

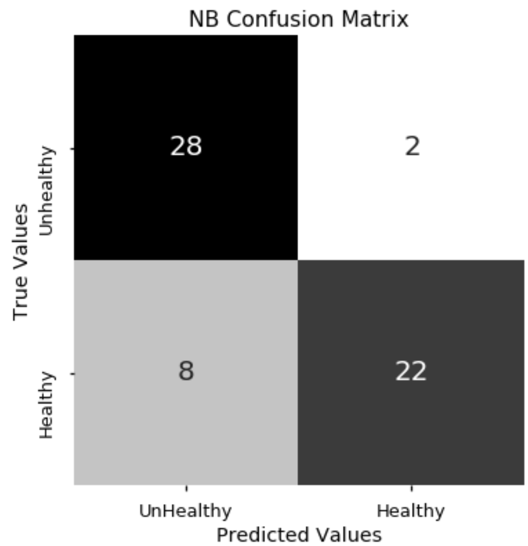


Figure 5-**1**: Mixed Naive Bayes confusion matrix

Figure 6-1 shows how Naïve Bayes was able to obtain an accuracy of 83.33%. Moreover, the achieved AUC of 0.872 shown in figure 6-2 explains how the naïve Bayes model performed in terms of avoiding false classifications.

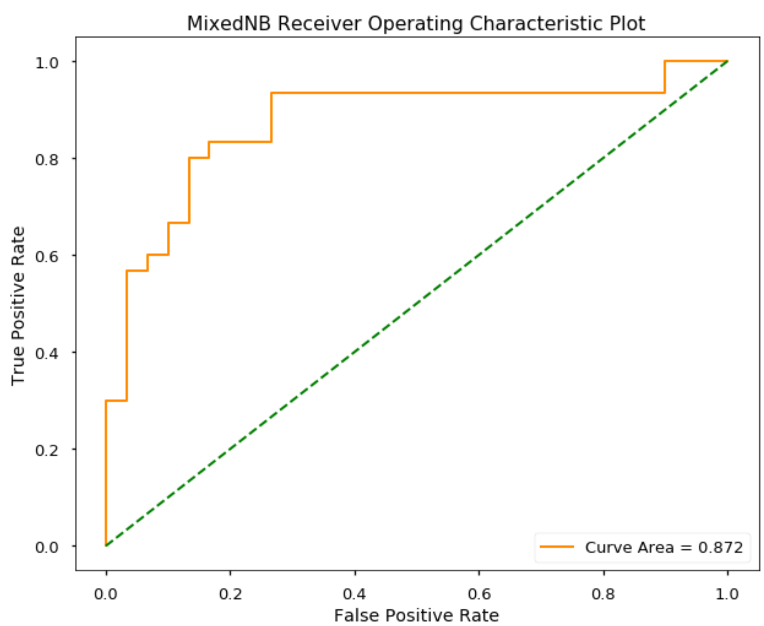


Figure 5-2: AUC of ROC curve for the mixed naive Bayes (MixedNB) model

Other measures such as precision, recall (sensitivity), and F-1 score are calculated per class. In precision, the MixedNB model obtained 0.78 and 0.92 for classes (0= unhealthy) and (1=healthy) respectively. For recall it was 0.93 and 0.73. finally, the F-1 score that measures how perfect are the precision and recall was 0.85 and 0.81.

**Logistic Regression**

Logistic Regression in Scikit-learn library in python support the feed of different parameters that could be used for optimization and regularization. Regularization is a technique used to solve the overfitting problem in machine learning models. The solver parameter selected was linear and C parameter, which indicates inverse of regularization strength. Different values were set for C and the selected one was 0.5.

Figures 6-3 and 6-4 show the confusion matrix and the AUC curve for the Logistic regression model.

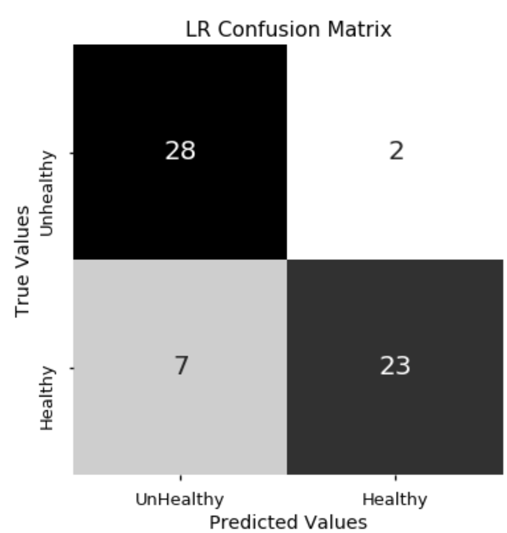


Figure 5-3: Logistic Regression Confusion Matrix

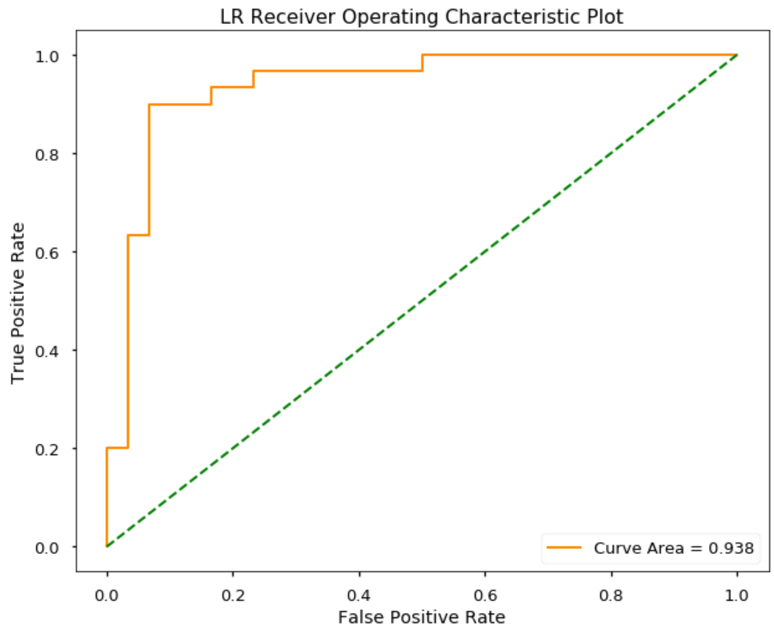


Figure 5-4: Logistic Regression AUC

Logistic regression achieved a minimal improvement in accuracy than NB where it obtained 0.85. For precision, LR got 0.8 and 0.92 for classes 0 and 1 and for recall it was 0.93 and 0.77. Finally, f1-score is 0.86 and 0.84 and an AUC of 0.938.

**Decision Trees**

Though the accuracy measure on the training dataset achieved using a 10-fold cross validation and the tuning of the different parameters is 0.84, decision tree suffered from the problem of overfitting and the accuracy dropped to 0.75 on the test set. Confusion matrix and AUC curve are shown in figures 5-5 and 5-6.

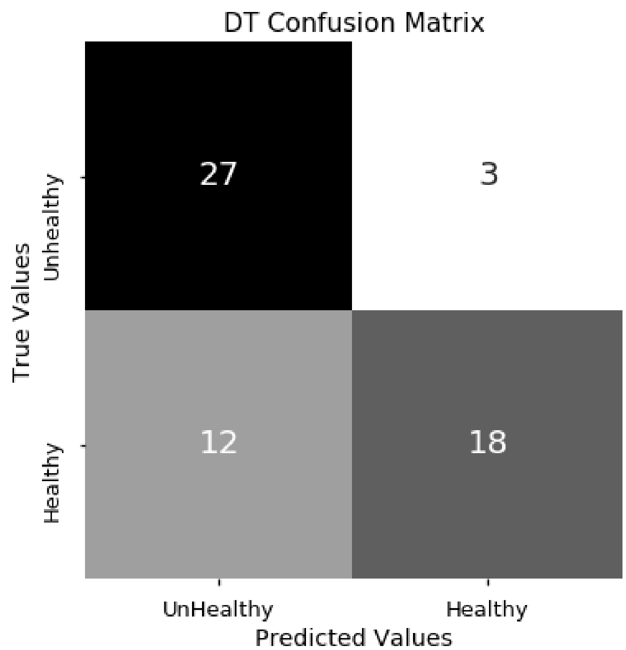


Figure 5-5: Decision Tree Confusion Matrix

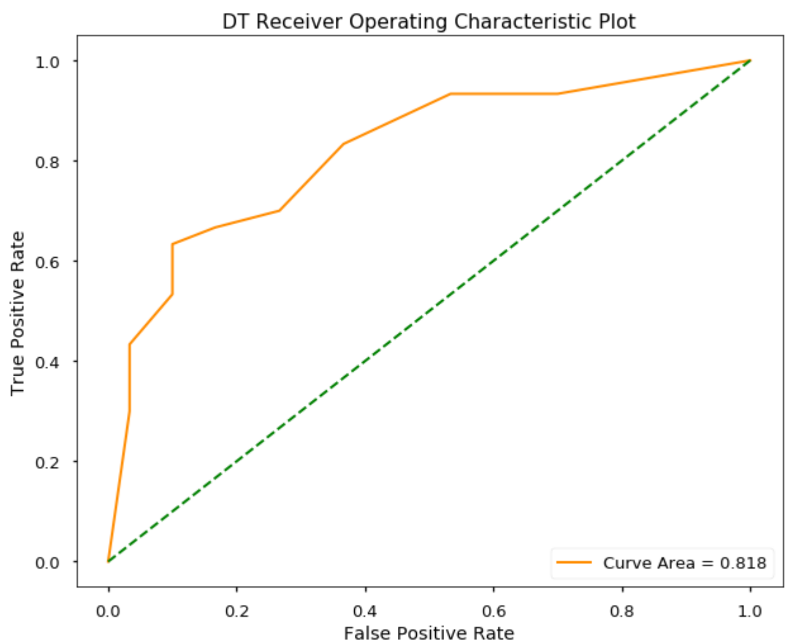


Figure 5-6: Decision Tree AUC

The decision tree suffered specifically in terms of sensitivity for the class of healthy people where it got a recall of 0.6. the f-1 score of the model was 0.78 for class 0 and 0.71 for class 1.

Same applications were conducted on SVM and ANN and the summary of all results are shown in table 5-1.

Table **5-1**: Summary of the measure for all the used models

| | | Precision | | Recall | | F1-Score | | |
|---|---|---|---|---|---|---|---|---|
| Model | Accuracy | Class 0 | Class 1 | Class 0 | Class 1 | Class 0 | Class 1 | AUC |
| NB | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 | 0.85 | 0.81 | 0.87 |
| LR | 0.85 | 0.80 | 0.92 | 0.93 | 0.77 | 0.86 | 0.84 | 0.94 |
| DT | 0.75 | 0.69 | 0.86 | 0.90 | 0.60 | 0.78 | 0.71 | 0.81 |
| SVM | 0.83 | 0.78 | 0.92 | 0.93 | 0.73 | 0.85 | 0.81 | 0.93 |
| ANN | 0.85 | 0.80 | 0.92 | 0.93 | 0.77 | 0.86 | 0.84 | 0.85 |

Table 5-1 summarizes the results obtained by the different methods. It is noted that all of the models performed somewhat similarly in terms of accuracy except the decision tree (DT) model which seem might suffer from the problem of overfitting or the using of many features. Logistic Regression model was the best models on all of the metrices achieving an accuracy of 85% and 94% AUC. Artificial neural network came second because of its AUC that was 85%.

## Model Explanation

This section will discuss the ways that can be used to explain and support the results of an AI model such as the logistic regression model in this case.

**Feature Importance**

In order to examine how the different features affect a certain model, a feature importance measure has to be calculated. This will lead to know which features affect the prediction of the model the most. One of the measures to calculate feature importance is permutation importance. Permutation feature importance is defined as the decrease in the model performance when a that feature column is randomly shuffled. Leo Breiman (2001) However, conclusions should be drawn on the model examined at hand without leading to insights about the intrinsic predictive power of the features themselves. Figure 5-7 shows the permutation feature importance measure for the variables in the LR model obtained using the eli5 library in python.

| Weight | Feature |
|---:|---|
| 0.0767 ± 0.0806 | ca |
| 0.0333 ± 0.0365 | thalach |
| 0.0267 ± 0.0267 | cp_2 |
| 0.0267 ± 0.0618 | oldpeak |
| 0.0200 ± 0.0249 | cp_3 |
| 0.0100 ± 0.0163 | trestbps |
| 0.0067 ± 0.0163 | fbs |
| 0.0067 ± 0.0340 | slope_2 |
| 0.0033 ± 0.0133 | slope_1 |
| 0 ± 0.0000 | restecg_2 |
| -0.0000 ± 0.0211 | exang |
| -0.0033 ± 0.0533 | thal_3 |
| -0.0067 ± 0.0163 | cp_1 |
| -0.0067 ± 0.0267 | chol |
| -0.0067 ± 0.0267 | age |
| -0.0100 ± 0.0400 | sex |
| -0.0100 ± 0.0340 | restecg_1 |
| -0.0167 ± 0.0211 | thal_2 |

Figure **5-7**: Mixed Naive Bayes confusion matrix

Figure 5-7 shows how much the accuracy of the LR model was affected by shuffling the columns of the different features. Positive or green values represent the most important features and how much decrease in accuracy occurred when the shuffle happened. The randomness of multiple shuffles is captured after the ± symbol. For example, ca or number of major vessels (0-3) colored by fluoroscopy is the most important feature for this model where shuffling it would decrease the model accuracy by up to 8%. On the other hand, negative values refer to the idea that shuffling those features resulted in an increase in the accuracy measure. This could have happened because of random chance only or because of how small the dataset is.

**LIME**

As explained earlier, Local Interpretable Model-agnostic Explanations (LIME) set by Ribeiro et al. (2016) aims to provide evidence for users to trust machine learning models. Trust that can be gained through explaining the predictions made by the model. The methodology of the approach is as follows:

1. Sampling instances close and far from the interpretable representation of the original input.
2. Calculating the prediction of these instances from their interpretable representation and builds a weighted linear model by minimizing the loss and complexity.

Notice that the samples weighting is based on the proximity from the original point and it decreases as the distance increases. Figure 5-8 shows an explanation from the original paper.

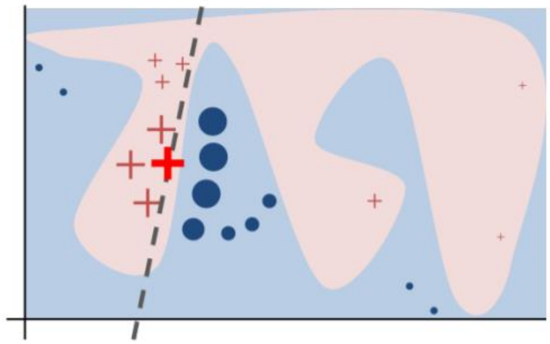

**Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function $f$ (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using $f$, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.**

Figure **5-8**: LIME methodology explained *taken as a copy from the original paper

One of limitations of this thesis is the lack of implementing LIME on logistic regression. Thus far, there is no available library to produce illustrative explanations using logistic regression models. However, both figures 5-9 and 5-10 show an application of how the LIME approach is done using the decision tree model. While figure 6-4 shows a correct prediction example, figure 6-5 shows an example of a misclassification.

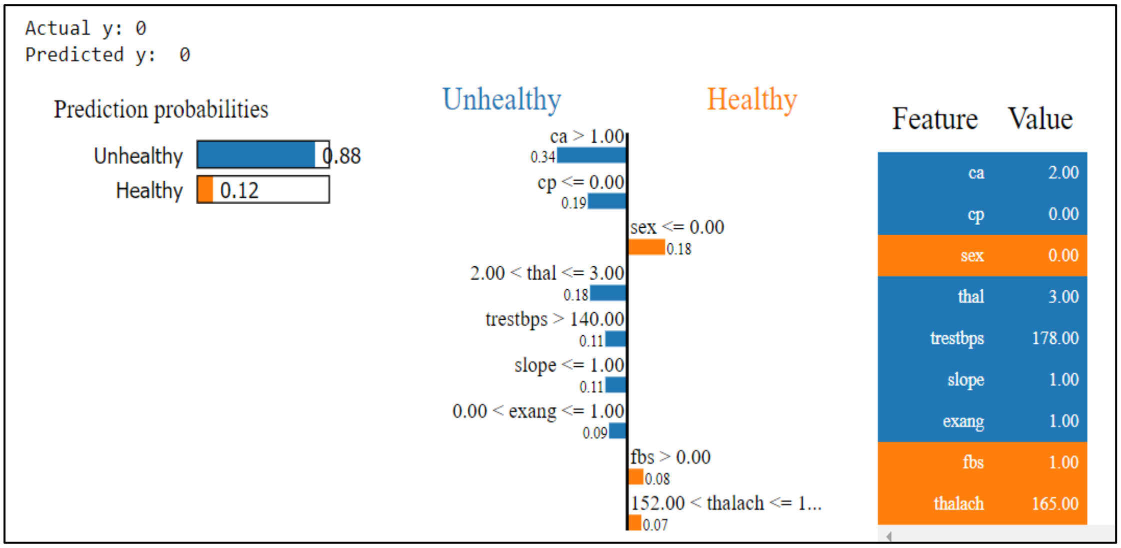


Figure **5-9**: Implementation of LIME with a correct prediction

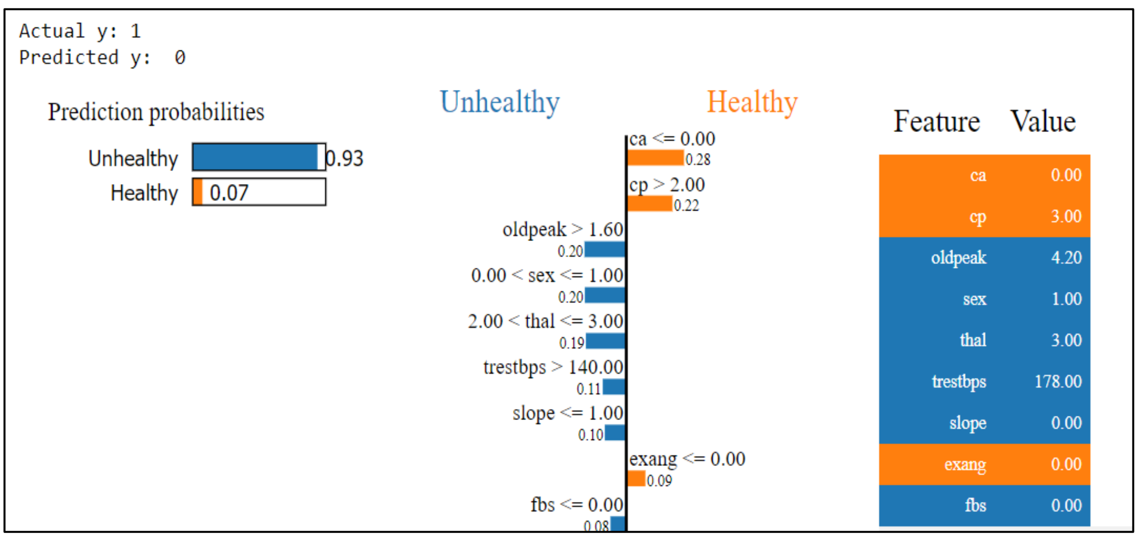


Figure **5-10**: Implementation of LIME with a wrong prediction

Figure 5-9 shows how the model weighted its prediction heavily on the value of ca=2, cp=0, and thal=3. On the other hand, figure 5-10 shows how the model established its misclassification depending on the values of oldpeak, sex, thal, and trestbps compared to the weights of ca and cp. Availability of such information gives the user, a doctor in this case, the choice of whether to trust the model or not.

**Chapter 6**

# Conclusions and Future Work

In this study, a literature survey has been conducted to provide an overview of the applications of AI and ML in healthcare sector. Various studies and published papers varied along the type of data used. Medical imaging, for example, is the richest body of literature when it comes to AI in healthcare. This is due to the progress and development in many of the computer vision algorithms represent by the deep learning convolutional neural network. Other studies utilized electrical signals and patient electronic records where algorithms of natural language processing are of much help for the latter.

Afterwards, a review on the applications of AI systems in cardiovascular diseases was presented with different studies utilizing the Cleveland dataset that is considered one of the most well-known datasets for applications of AI and ML in cardiovascular medicine and is open for nonspecialized researchers. Accompanying this was the study of Ribeiro et al. (2016) that is of core interest of this paper asserting the importance of building trust in AI models.

From there, the Cleveland dataset was explained and processed to be prepared for different classification algorithms. Namely, Naïve Bayes, Logistic Regression, Decision Trees, Support Vector Machine, and Artificial Neural Network. Later on, metrics of explaining the best model were measured and explained.

The key findings of this study are as follow:

1. Logistic Regression achieved the best area under the curve result; however, it was in a tie with artificial neural network in all other metrics such as recall and accuracy.
2. Decision Trees suffered from the overfitting problem by using all the features though measures of tuning the best parameters was carried.
3. Number of major vessels colored by fluoroscopy, or Ca, was the most important feature that affect the accuracy of the Logistic Regression model. Randomly shuffling the column change the accuracy of the model by a value up to 7%.

The results of this study had some shortcomings that could be handled in future work. For example, the implementation of LIME technique was conducted on a different model. It was made on a decision tree model rather than a logistic regression model to explain the process due to lack of a specific library that can carry LIME method for a Logistic Regression model. Another shortcoming is the size of the dataset where there are only 303 instances with some missing values. Future work could benefit from incorporating different datasets to improve the training results.

Moreover, though measures to explain the model was carried, there is still a need for a more scrutinized, auditing, and validation before such a system is implemented in the hospitals since the results can be catastrophic. Finally, deep neural networks applications can seem promising but conducting applications requires open access to datasets that still not available to the out of the field scientists. A solution could be to enable access after ensuring the appropriate handling of privacy measures.

# References


Abràmoff, M.D., Lavin, P.T., Birch, M. et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Med 1, 39 (2018) doi:10.1038/s41746-018-0040-6

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adeli, H. (2018). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine*, *100*, 270–278. doi: 10.1016/j.compbiomed.2017.09.017

Acharya, U. R., Fujita, H., Oh, S. L., Raghavendra, U., Tan, J. H., Adam, M., … Hagiwara, Y. (2018). Automated identification of shockable and non-shockable life-threatening ventricular arrhythmias using convolutional neural network. *Future Generation Computer Systems*, *79*, 952–959. doi: 10.1016/j.future.2017.08.039

Afzal, N., Sohn, S., Abram, S., Scott, C. G., Chaudhry, R., Liu, H., … Arruda-Olson, A. M. (2017). Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. *Journal of Vascular Surgery*, *65*(6), 1753–1761. doi: 10.1016/j.jvs.2016.11.031

Atzori, M., Cognolato, M., & Müller, H. (2016). Deep Learning with Convolutional Neural Networks Applied to Electromyography Data: A Resource for the Classification of Movements for Prosthetic Hands. *Frontiers in Neurorobotics*, *10*. doi: 10.3389/fnbot.2016.00009

Bote-Curiel, L., Muñoz-Romero, S., Gerrero-Curieses, A., & Rojo-Álvarez, J. L. (2019). Deep Learning and Big Data in Healthcare: A Double Review for Critical Beginners. Applied Sciences, 9(11), 2331. doi: 10.3390/app9112331

Breiman, L. Machine Learning (2001) 45: 5. https://doi.org/10.1023/A:1010933404324

Cheng, Y., Wang, F., Zhang, P., & Hu, J. (2016). Risk Prediction with Electronic Health Records: A Deep Learning Approach. *SDM*.

Cleveland Heart Disease Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart Disease.

D.E. Rumelhart, G.E. Hinton, R. J. Williams, ―Learning representations by back-propagating errors,‖ Nature 323 (6188), pp. 533-536, 1986.

Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert Systems with Applications*, *36*(4), 7675–7680. doi: 10.1016/j.eswa.2008.09.013

Du, L.-H., Liu, W., Zheng, W.-L., & Lu, B.-L. (2017). Detecting driving fatigue with multimodal deep learning. *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*. doi: 10.1109/ner.2017.8008295

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., . . . and the CAMELYON16 Consortium. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama, 318(22), 2199-2210. doi:10.1001/jama.2017.14585

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. doi:10.1038/nature21056

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, *161*, 1–13. doi: 10.1016/j.cmpb.2018.04.005

Fraiwan, L., & Lweesy, K. (2017). Neonatal sleep state identification using deep learning autoencoders. *2017 IEEE 13th International Colloquium on Signal Processing & Its Applications (CSPA)*. doi: 10.1109/cspa.2017.8064956

Geng, W., Du, Y., Jin, W., Wei, W., Hu, Y., & Li, J. (2016). Gesture recognition by instantaneous surface EMG images. *Scientific Reports*, *6*(1). doi: 10.1038/srep36571

Gudadhe, M., Wankhade, K., & Dongre, S. (2010). Decision support system for heart disease based on support vector machine and Artificial Neural Network. *2010 International Conference on Computer and Communication Technology (ICCCT)*. doi: 10.1109/iccct.2010.5640377

Jaymin Patel, Prof.TejalUpadhyay and Dr. Samir Patel, "Heart Disease Prediction using Machine Learning and Data Mining Techniques ", Nirma University, Gujarat, India IJCSC Vol 7,number 1 september 2015-march 2016 pp.129- 137.

Jiang F, Jiang Y, Zhi H, *et al*. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology* 2017;2: e000101. doi:10.1136/svn-2017-000101

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications*, *35*(1-2), 82–89. doi: 10.1016/j.eswa.2007.06.004

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.-J., & Fei-Fei, L. (2018). Thoracic Disease Identification and Localization with Limited Supervision. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. doi: 10.1109/cvpr.2018.00865

Majumdar, A., & Ward, R. (2017). Robust greedy deep dictionary learning for ECG arrhythmia classification. *2017 International Joint Conference on Neural Networks (IJCNN)*. doi: 10.1109/ijcnn.2017.7966413

Medhekar, D.S., Bote, M.P., Deshmukh, S.D.: Heart disease prediction system using naive bayes. Int. J. Enhanced Res. Sci. Technol. Eng. **2**(3), (2013)

M.T. Ribeiro, S. Singh, C. Guestrin (2016). **“**Why should I trust you?”: explaining the predictions of any classifier *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1135-1144

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2016). DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. *Advances in Knowledge Discovery and Data Mining Lecture Notes in Computer Science*, 30–41. doi: 10.1007/978-3-319-31750-2_3

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. doi: 10.1038/s41591-018-0300-7

Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015) Heart Diseases Detection Using Naive Bayes Algorithm. IJISET-International Journal of Innovative Science, Engineering & Technology, 2, 441-444

Vine, L. D., Zuccon, G., Koopman, B., Sitbon, L., & Bruza, P. (2014). Medical Semantic Similarity with a Neural Language Model. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM 14*. doi: 10.1145/2661829.2661974

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2097–2106. doi: 10.1109/cvpr.2017.369

Xia, B., Li, Q., Jia, J., Wang, J., Chaudhary, U., Ramos-Murguialday, A., & Birbaumer, N. (2015). Electrooculogram based sleep stage classification using deep belief network. *2015 International Joint Conference on Neural Networks (IJCNN)*. doi: 10.1109/ijcnn.2015.7280775

Xia, P., Hu, J., & Peng, Y. (2017). EMG-Based Estimation of Limb Movement Using Deep Learning With Recurrent Convolutional Neural Networks. *Artificial Organs*, *42*(5). doi: 10.1111/aor.13004

Zhang, N., Yang, G., Gao, Z., Xu, C., Zhang, Y., Shi, R., … Firmin, D. (2019). Deep Learning for Diagnosis of Chronic Myocardial Infarction on Nonenhanced Cardiac Cine MRI. *Radiology*, *291*(3), 606–617. doi: 10.1148/radiol.2019182304

Zheng, W.-L., Zhu, J.-Y., Peng, Y., & Lu, B.-L. (2014). EEG-based emotion classification using deep belief networks. *2014 IEEE International Conference on Multimedia and Expo (ICME)*. doi: 10.1109/icme.2014.6890166

Zheng, Y., Liu, Q., Chen, E., Ge, Y., & Zhao, J. L. (2014). Time Series Classification Using Multi-Channels Deep Convolutional Neural Networks. *Web-Age Information Management Lecture Notes in Computer Science*, 298–310. doi: 10.1007/978-3-319-08010-9_33

Zhou, Z., & Jiang, Y. (2004). NeC4.5: Neural ensemble based C4.5. *IEEE Transactions on Knowledge and Data Engineering, 16*(6), 770-773. doi:10.1109/TKDE.2004.11

Zhu, X., Zheng, W.-L., Lu, B.-L., Chen, X., Chen, S., & Wang, C. (2014). EOG-based drowsiness detection using convolutional neural networks. *2014 International Joint Conference on Neural Networks (IJCNN)*. doi: 10.1109/ijcnn.2014.6889642