T.C.
ISTANBUL AYDIN UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

CARDIOVASCULAR DISEASES DETECTION USING ARTIFICIAL INTELLIGENCE

MASTER'S THESIS

Ayodele Martin DOSSOU

Department of Software Engineering
Artificial Intelligence and Data Science Program

JUNE, 2023

T.C.
ISTANBUL AYDIN UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

CARDIOVASCULAR DISEASES DETECTION USING ARTIFICIAL INTELLIGENCE

MASTER'S THESIS

Ayodele Martin DOSSOU
(Y1913.140001)

Department of Software Engineering
Artificial Intelligence and Data Science Program

Thesis Advisor: Prof. Dr. Ali OKATAN

JUNE, 2023

APPROVAL PAGE

DECLARATION

I hereby declare with respect that the study “Cardıovascular Dıseases Detectıon Usıng Artıfıcıal Intellıgence”, which I submitted as a Master thesis, is written without any assistance in violation of scientific ethics and traditions in all the processes from the project phase to the conclusion of the thesis and that the works I have benefited are from those shown in the Bibliography. (0/06/2023)

Ayodele Martin DOSSOU

FOREWORD

It has been written to fulfill the graduation requirements of the Master in Artificial Intelligence and Data Science at Istanbul Aydin University. I would like to thank Dr. ALI OKETUN for her guidance and support throughout this process. I also wish to thank all of the respondents; without whose cooperation I would not have been able to conduct this analysis. Thanks also to the members of the committee who attended my master's thesis defense. I appreciated the chance to discuss issues with my friends and family. It kept me motivated whenever I lost interest. I particularly admire the wisdom and kindness of my parents: they have always given me invaluable advice and support.

January 2023 Ayodele Martin DOSSOU

CARDIOVASCULAR DISEASES DETECTION USING ARTIFICIAL INTELLIGENCE

ABSTRACT

The likelihood of contracting a disease rises with the size of the human population. Globally, there are numerous ailments, and one of the main issues facing Healthcare systems now lack the required technology to detect illness in patients. Cardiovascular disease, or CVD, is one such illness. Any cardiovascular, vascular, or blood vessel ailment is referred to. More people globally die from CVDs than from the WHO. More so in low- and middle-income nations. When ill, it can be quite difficult for any other cause, according for persons who live alone to contact the hospital. As a result, we created a simulation that is capable of when A sick patient notifies the hospital in writing. a simulation that is capable of. Currently, the simulation merely detects and informs the hospital about patients with cardiovascular disease. We chose to focus on heart disease detection because it's one of the worst diseases and there's a significant chance that people may pass away from it. It is a classification problem to determine if a patient has heart disease or not. Age, blood sugar, cholesterol, and many other factors are considered, and the output is then provided based on the input.

We leverage both traditional machine learning and state-of-the-art deep learning techniques. The machine learning techniques include a support vector machine (SVM) with Artificial Neural Network (ANN) , logistic Regression , and

Keywords: Feature selection, Support Vector Machine (SVM), Transfer Function, Artificial Neural Network (ANN), BMI

YAPAY ZEKA İLE KARDİYOVASKÜLER HASTALIKLAR TESPİTİ

ÖZET

Bir hastalığa yakalanma olasılığı, insan popülasyonunun büyüklüğü ile birlikte artar. Küresel olarak, çok sayıda hastalık var ve Sağlık sistemlerinin karşı karşıya olduğu ana sorunlardan biri artık hastalardaki hastalığı tespit etmek için gerekli teknolojiden yoksun. Kardiyovasküler hastalık veya CVD, böyle bir hastalıktır. Herhangi bir kardiyovasküler, vasküler veya kan damarı rahatsızlığına atıfta bulunulur. Küresel olarak KVH'lerden ölen insan sayısı DSÖ'den ölenden daha fazladır. Daha çok düşük ve orta gelirli ülkelerde. Hastayken başka bir nedenle hastaneye başvurmak yalnız yaşayan kişilere göre oldukça zor olabilir. Sonuç olarak, hasta olan bir hastayı hastaneye yazılı olarak bildirdiğinde bunu yapabilen bir simülasyon oluşturduk. yapabilen bir simülasyon. Simülasyon şu anda yalnızca kardiyovasküler hastalığı olan hastaları tespit ediyor ve hastaneyi bilgilendiriyor. Kalp hastalığı tespitine odaklanmayı seçtik çünkü bu en kötü hastalıklardan biridir ve insanların ondan ölme olasılığı yüksektir. Bir hastanın kalp hastası olup olmadığının belirlenmesi bir sınıflandırma problemidir. Yaş, kan şekeri, kolesterol ve diğer birçok faktör dikkate alınır ve ardından girdiye göre çıktı sağlanır.

Hem geleneksel makine öğreniminden hem de son teknoloji ürünü derin öğrenme tekniklerinden yararlanıyoruz. Makine öğrenimi teknikleri arasında Yapay Sinir Ağı (YSA) ile bir destek vektör makinesi (SVM), lojistik Regresyon,

Anahtar Kelimeler: Özellik seçimi, Destek Vektör Makinesi (SVM), Transfer Fonksiyonu, Yapay Sinir Ağı (YSA), BMI

TABLE OF CONTENT

ABBREVIATIONS

AUC : Area Under the Curve

CVD : Cardiovascular disease

WHO : World Health Organization

AAN : Artificial Neural Network

AI : Artificial Intelligence

SVM : Support Vector Machine

BMI : Body Mass Index

CNN : Convolution Neural Network

ECG : Electrocardiogram

NB : Naves Bayes

AHA : American Heart Association

MLP : Multi-layer Neural Network

TP : True Positive

DT : Decision Tree

TN : True Negatives

FN : False Negatives

FS : Feature Selection

FP : False positive

KNN : K-Nearest Neighbor

NB : Naïve Bayes

SVM : Support Vector Machine

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

The research primarily focuses on the numerous category segregation techniques used to forecast cardiac disorders. A poor lifestyle, drinking alcohol, eating a lot of fat, triggering hypertension, and not getting enough exercise all contribute to heart disease. The heart of a human controls blood flow throughout the body. The majority of the human body is made up of it. Heart irregularities are a severe cause for concern because they have an impact on many different body organs. Heart illness can be thought of as irregularities or abnormalities in how the heart usually beats.

Heart congestion and disease are the primary causes of the majority of fatalities in today's fast and busy environment. [1] According to the WHO, more than 10 million people worldwide die from heart disease each year. The only strategies to stop heart-related illnesses are through early identification and a healthy lifestyle. Today's world makes it exceedingly challenging for medical centers to provide individualized providing for each patient's needs because of the growing population. In light of this, When a patient becomes unwell, we have made the decision to attempt to alert the hospitals. The project generally concentrates on heart illness [2] as a result of a very high possibility that a patient would suffer a serious injury or pass away as a result of it. Additionally, it shortens the time it takes to detect cardiac disease.

This initiative is crucial because, if a patient is living alone and suffering from heart illness, he may not be able to ask the hospital staff for assistance. Our effort goes a long way towards assisting such people. We faced difficulties and hurdles because we lacked sufficient data sets to produce pleasing results, which we overcame by producing synthetic data. Preparing the method to decide if the patient is unwell or not, we employ machine learning modules. Even if our experiments produced solid results, we still need to put the system into practice, It entails developing a mobile application that gathers patient data and transmits it to hospitals. This could eventually be put into practice.

II. LITERATURE REVIEW

Deep learning-based cardiovascular detection has received a lot of attention recently. In order to determine the viability and precision of employing deep learning models for cardiovascular disease detection, several research investigations have been carried out. The following are some related works in this field:

- "Cardiovascular Disease Detection Using Deep Learning: A Review" [4] by Liang et al. from 2020 states that this work provides a complete analysis of recent advancements in deep learning-based cardiovascular disease diagnosis. The authors discuss the many sorts of cardiovascular diseases as well as the deep learning techniques used to diagnose them. They also provide an overview of the datasets used in these studies as well as the performance standards that were used to judge the performance of the models.
- Convolutional neural networks (CNNs) were used in the 2019 study "Automated Detection of Cardiovascular Disease Using Deep Learning Techniques" by Attia et al. to classify patients as having cardiovascular disease or not based on the images from their echocardiograms. The scientists' approach had an accuracy of 80.8% in detecting cardiovascular disease, outperforming traditional machine learning algorithms.
- "Cardiovascular Disease Prediction using Deep Learning Algorithm" by Gupta et al. (2020): In this work [6], the authors employed a deep learning model to forecast cardiovascular illness based on demographic, lifestyle, and medical history information. The accuracy of the authors' prediction of cardiovascular disease was 88.23%, exceeding the performance of conventional machine learning models.
- " In Singh et al.'s article "Deep Learning-based Detection of Cardiac Arrest Using ECG Signals" from 2021, they created the following: This work [7] built a deep learning model to identify cardiac arrest using electrocardiogram (ECG) information. In comparison to standard machine-learning techniques,

the authors' method had a 98.6% accuracy rate for detecting cardiac arrest.

- Xue, J. et al., "Machine learning for cardiovascular disease detection and diagnosis." It's important to recognize and diagnose cardiovascular illness. [8] In this study, the potential use of **electrocardiogram (ECG)** and medical imaging signals to provide patient data to machine learning algorithms is investigated.
- "Biomarkers in cardiovascular disease: Prospects for personalized diagnosis and treatment," M. V. Kamath et al. [9] This study examines the use of biomarkers for cardiovascular disease risk assessment, early identification, and individualized management.
- "Cardiac Magnetic Resonance Imaging in the Detection of Cardiovascular Disease," by S. K. White et al., says that.[10] In this study, the effectiveness of cardiac magnetic resonance imaging (MRI) as a non-invasive diagnostic technique is assessed in order to identify and diagnose cardiovascular disease.
- "Mobile health applications for the detection and management of cardiovascular disease." Y. Wang et al. This review article explores the [11] use of mobile health applications to identify and treat cardiovascular disease by keeping track of symptoms and vital signs, monitoring medication compliance, and providing personalized treatment recommendations.
- " The study by M. S. Khan et al., "Early detection of cardiovascular disease using machine learning techniques on electronic health records," details a thorough investigation of the topic.[12] This systematic study examines the effectiveness of machine learning algorithms in identifying cardiovascular sickness using electronic health information, and it highlights the potential for improved early identification and tailored therapy.
- Theresa Prince, R., et al. (2016) conducted a review of the various heart disease prediction models. Theresa employed Naive Bayes and Neural Networks as her categorization methods.

- the LR, **the KNN network**, and the decision tree. Success was had in comparing the accuracy ratings for each model [3].
- Overall, these findings highlight the need for additional research in this area and demonstrate the promise of deep learning models for detecting cardiovascular illness.

III. PROPOSED SCHEME AND DATASET DETAILS

A. Dataset Analyzation

The study aims to detect and diagnose cardiovascular disease using datasets and utilizing many classification methods.

- **Cardiovascular Disease Dataset**

Using the cardiovascular disease dataset was identified and identified using this inquiry, and the findings were compared to those of previous investigations. It has a lot of patient information, including medical records. Kaggle's dataset was gathered from three sources, and they are Examining the findings of several medical tests, Objective reflects the information provided by the patient, and Subjective represents the data gathered as facts about cardiovascular illnesses. the set of data used for training, testing, and validation. The website contains the intended data, which is open to the public [10].

The shape of the cardiovascular disease's dataset is (68783, 12), and it is a clean version of the CVD dataset as follows:

Table 1: cardiovascular diseases dataset

Features	Descriptions
Age	The patient age in years
Gender	Gender of patient (1: Male, 0: Female)
Height	Representing the height of patient's
Weight	Representing the weight of patient's
Systolic BP	Systolic blood pressure
Diastolic BP	Diastolic blood pressure
Cholesterol	The Cholesterol Level in the blood (1: normal, 2: above normal, 3: well above normal)
Glucose	Categorical value of the sugar blood level (1: normal, 2: above normal, 3: well above normal)
Smoke	Smoking (0: No, 1: Yes)
Alcohol	Alcohol intake (0: No, 1: Yes)
Physical Activity	Physical activity type
Cardio Disease	Target value measuring the Presence or absence of cardiovascular disease.

Table 2: Class Distribution CVD

Class	Counts
0	35021
1	34979

35,021 out of 70,000 cases in this dataset are labelled as having no cardiovascular disease, and 34,979 cases are labelled as having cardiovascular disease. This suggests that the dataset is roughly balanced.

B. Exploratory data analysis:

Initial data analysis entails looking for trends and discrepancies in data using summary statistics and graphical representation.

1. Eliminating Duplicate and Unnecessary Data:

We eliminate all of the duplicate and Nan values from the data set. We see that the dataset has certain inconsistencies, such as the minimum age of 29 years and the minimum weight of 10 kg. In several circumstances, the systolic blood pressure was greater than the diastolic blood pressure. To remedy the inaccuracies, we, therefore, eliminate the outliers. Outliers are data mistakes that have the potential to seriously skew the outcome. After cleaning the data, No datasets exist, as shown by the box plot in the accompanying graph, where the systolic pressure is higher than the diastolic pressure. The dataset we used didn't have any Null values, however it did include duplicate entries, as we found.

We remove the 24 duplicate values from the sample as a result, leaving us with a dataset containing 60118 data points.

2. Correlation:

All of the features' correlation matrices, which depict how strongly one feature is connected with another, have been plotted. By doing this, we can identify any features that can skew the results and get rid of them. We can see from the correlation matrix below that age and cholesterol have a significant impact on the result.

A heatmap shows correlations between variables in the dataset. From the above heatmap, we see that no features directly correlate with the label (cardio) and that generally, there do not exist significant correlations between other variables. However, correlations exist between a select few variables:

- Height and gender are correlated with a score of 0.5.
- Glucose and cholesterol are correlated with a score of 0.45.
- Smoking habits and gender are correlated with a score of 0.34

Figure. 1: Correlation Matrix prior to Feature Engineering

3. check outliers

The above boxplots confirm our suspicion that there are outliers in the dataset. There appear to be several outliers in the systolic and diastolic variables (ap_hi and ap_lo). These outliers might be explained by human error when entering data into the .csv format. Our prediction model might benefit if these outliers were removed from the data.

```
# check outliers

def check_outliers(df):
    l = df.columns.values
    number_of_columns=(len(l)-1)/2
    number_of_rows = 2

    plt.figure(figsize=(4*number_of_columns,8*number_of_rows))
    for i in range(1,len(l)):
        plt.subplot(number_of_rows + 1,number_of_columns,i)
        sns.set_style('whitegrid')
        sns.boxplot(df[l[i]],orient='v')
        plt.tight_layout()

check_outliers(df)
```

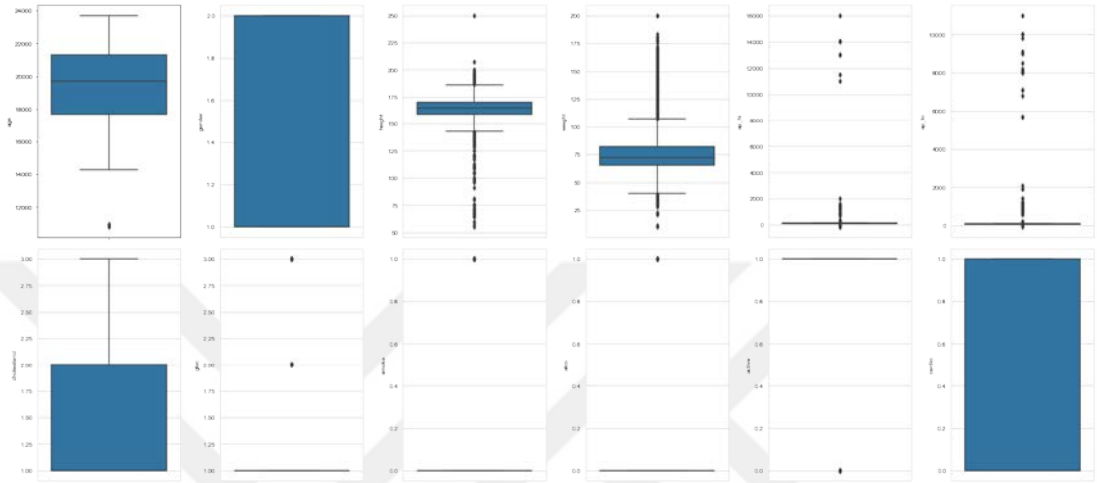


Figure 2: Box Plot confirming the presence of of Outliers

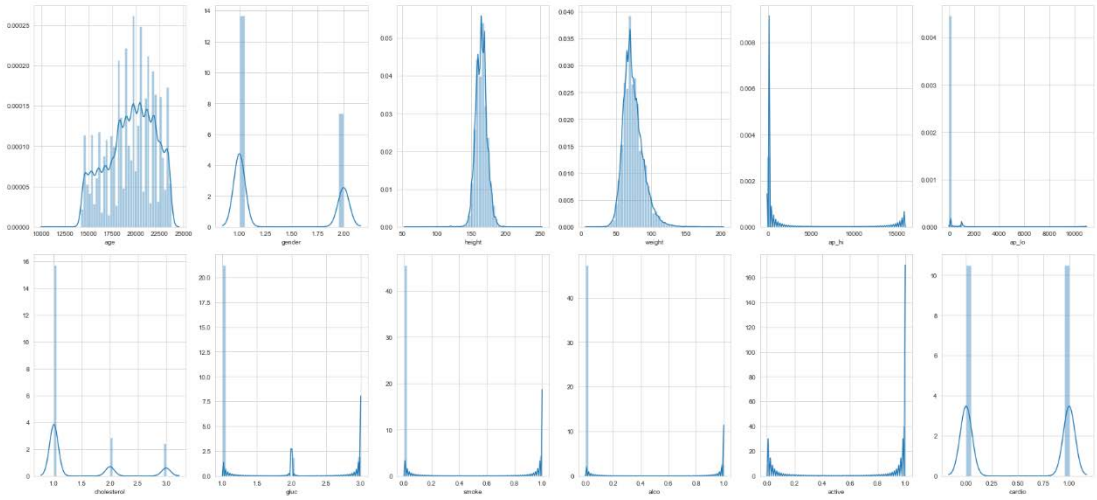
4. check distribution and skew

Histograms show the distribution of each feature in the dataset. There appears to be some variation in age, but height and weight are roughly normally distributed. However, the histograms further suggest the presence of outliers in the blood pressure features. Moreover, there exist categorical features in the dataset, such as gender, cholesterol, glucose, smoking habits, alcohol use, and activity.

```
# check distribution and skew
def check_dist(df):
    l = df.columns.values
    number_of_columns=(len(l)-1)/2
    number_of_rows = 2

    plt.figure(figsize=(4*number_of_columns,8*number_of_rows))
    for i in range(1,len(l)):
        plt.subplot(number_of_rows+1,number_of_columns,i)
        sns.distplot(df[l[i]],kde=True)
        plt.tight_layout()

check_dist(df)
```



Fiureg 3: Histograms show the distribution of each features

The dataset was preprocessed to remove outliers in the ap_hi and ap_lo blood pressure features. Only values less than 250 were kept in the data. Through this process, almost **1000 datapoints** were discarded, but the resultant distributions depicted in the above bar plots and histograms for ap_hi and ap_lo are roughly normal.

Note that we are dealing with a mix of continuous, ordinal, and binary data, as confirmed from the above data analysis. We can combine all of these data types in one model. Here are a few typical steps we took to preprocess this data before modeling.

- Standardize all continuous features: All continuous input should be standardized. For every continuous feature, compute its mean (μ) and standard deviation (σ) and calculate $\boldsymbol{x = (x-\mu) / \sigma}$.
- Binarize categorical/discrete features (create dummy variables): For all categorical features, represent them as multiple boolean features. For example, instead of having one feature called cholesterol, have 3 boolean features - chol_normal, chol_above_normal, chol_well_above_normal and appropriately set these features to 0 or 1. As can be seen, for every categorical feature, k binary features are added, where k is the number of values that the categorical feature takes.

5. Feature Engineering:

The Body Mass Index **(BMI)** was determined during the data processing phase to assess the patient's health. A normal **BMI** is between 18 and 25, while an unhealthy **BMI** is over 25 or under 18. The formula below is used to compute the **BMI** values. The features for height and weight were then eliminated.

$$BMI = \frac{\text{Weight (KG)}}{\text{Height (cm)}}$$

The American Heart Association (AHA) claims that the systolic and diastolic blood pressure readings.

There are five gradations of severity for it. Each record's blood pressure was calculated, and the level was indicated.

6. Feature Selection:

Based on the features' relevance, we choose them [13]. The models' accuracy is unaffected by the pertinent features. As a result, we use a correlation matrix to choose relevant data. Gender is the factor that is least connected with the target, according to the correlation matrix, which also shows that features like BMI, weight, glucose, height, smoking, alcohol use, and activity level do not have high correlations with the target. As a result, we eliminate characteristics like BMI, weight, glucose, gender, height, and alcohol and/or drug use.

7. Feature Scaling:

The entire feature set in the data set can be normalized using this technique [14]. The model frequently has a tendency to favor larger values when we have a characteristic with a very high value. For the feature scaling, we used the Standardization formula.

8. Study Population

Participants were chosen using a multistage (prefecture-county-township-village) stratified cluster random sampling method. First, we selected Yili as the prefecture that best represented the Kazakh population in Xinjiang. Second, we chose

one township from each county and one county within each prefecture at random. Finally, the associated villages in each township were chosen using a stratified sampling technique. This study's prospective cohort was done in the Xinjiang Kazakh Autonomous Region's Nalati town. Between 2009 and 2013, a total of 1771 local Kazakh Chinese subjects under the age of 18 who had lived in the hamlet for at least six months were successfully enrolled. By the end of 2016, 1508 of them had complete information.

Before the baseline survey, those who had a history of CVD were not included. Before joining the study, each subject gave their signed, informed consent. The study was authorized by the Institutional Ethics Review Board of the First Affiliated Hospital of Shihezi University (IERB number. SHZ2010LL01).

9. Assessment of Variables

For the analysis in this study, we gathered 31 potential variables, including sociodemographic traits, medical history, dietary preferences, lab tests, and artificial indices. Professionally trained individuals took anthropometric measurements such as height, weight, waist, hip, and blood pressure. After a 5-min sat rest, blood pressure was taken three times in each individual, and the average reading was calculated. Each patient gave a sample of 5-mL fasting blood. Participants self-reported their current alcohol consumption and cigarette smoking habits. Similar to the definition of a family history of hypertension, a family history of diabetes is one in which at least one parent or sibling has had the disease. Systolic blood pressure (SBP) greater than 140 mmHg, diastolic blood pressure (DBP) below 90 mmHg, or use of antihypertensive drugs were all considered to be hypertension. In the Biochemistry Laboratory of the First Affiliated Hospital of Shihezi University School of Medicine, the fasting blood glucose (FBG), low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL), total cholesterol (TC), and triglycerides (TG) were assessed using a modified hexokinase enzymatic method. According to the China Adult Dyslipidemia Prevention Guide (2007) and IDF diagnostic criteria, respectively, the terms metabolic syndrome and dyslipidemia were defined. Using kits from Randox Laboratories Ltd. (Shanghai, China) and Elabscience Biotechnology (Wuhan), cytokines such as nonesterified fatty acids (NEFAs), high-sensitivity C-reactive protein (hs-CRP), adiponectin (ADP), insulin

(INS), and interleukin-6 (IL-6) were identified.

In addition, we calculated several artificial indices, such as the body mass index (BMI), the body adiposity index (BAI), the waist-to-hip ratio (WHtR), the waist-to-height ratio, the TGHR, the TCHR, the LDL/HDL ratio, and the MAP (mean arterial pressure, (DBP)*(2/3)+ (SBP)*(1/3)). This study followed the same methodology as our earlier research, and the description of the procedures partially uses their phrasing.21

10. CVD Event Ascertainment

The first known diagnosis of CVD was the main result of the analysis in this study. An ischemic heart disease, cerebrovascular disease, or any associated disease-related hospitalization or mortality was referred to as a CVD event (ICD9: Codes 390–495). Medical data from a nearby hospital, insurance claims, survey replies, death registries from the morbidity and mortality monitoring system, and survey responses from the follow-up period were used to identify CVD events. We followed up twice, once in 2012 and once in 2016.

During a face-to-face meeting, trained investigators collected the questionnaire replies. In November, we usually checked in with the subjects. In the questionnaire, we would first note the fundamental demographic data and the duration of the follow-up. If the subject passed away during the follow-up period, their relatives were questioned about the subject's passing, including the time, location, and cause of death. This information was then cross-referenced with data from the cause of death monitoring system. If the subjects lived, they would be questioned about whether they had ever been hospitalized as well as the cause(s) and length of stay, and the information would then be checked against medical insurance and record information to determine the hospitalization diagnosis.

11. Statistical Analysis,

We used the test set to provide the discrimination, calibration, and clinical utility of each ML model for model comparison. AUC was utilized to measure discrimination, and the DeLong test31 was used to compare the AUC of each ML model. The highest Youden index, which optimizes the sum of specificity and sensitivity, was

used to establish the best threshold probability for each model to identify high-risk participants. We also reported additional diagnostic test parameters below the ideal threshold, such as specificity, sensitivity, negative predictive value (NPV), and positive predictive value (PPV). Brier score 32 and the calibration curve were used to evaluate the calibration. The Brier score's confidence interval was computed using a 1000-times bootstrap. A score of 20 or a P-value of >0.05 in the Hosmer-Lemeshow chi-square statistic (2) denotes excellent calibration.33

The decision curve analysis (DCA)34 was used to evaluate the clinical applicability of the best-performing model, which was chosen by a combination of calibration and discrimination.

When necessary, the Student's t-test or the Mann-Whitney test for continuous variables and chi-square tests for categorical variables were used to compare baseline features. According to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD), we present our findings.35 We used R version 3.3 (http://www.r-project.org, The R Foundation) and scikit-learn in Python version 3.7 (Python Software Foundation) to carry out all statistical analyses. Statistical significance was defined as a 2-sided P value 0.05.

IV. PROPOSED APPROCH:

The Gradient Boosting, Decision Tree (DT), Logistic Regression, and Neural Network algorithms will all be used in this study to classify data (NB). In order to discover the best classifier, it is also necessary to create a deep network and assess the effects of different optimization learning techniques on the detection of cardiovascular illnesses.

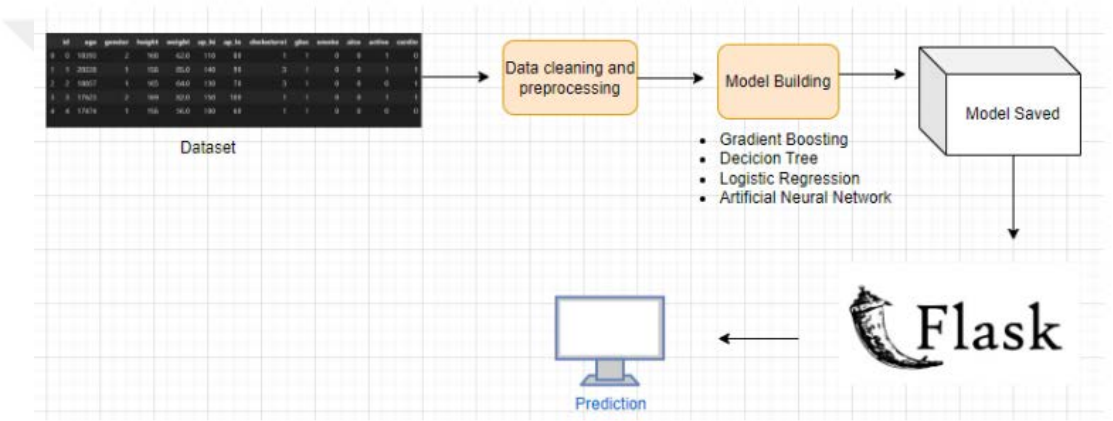


Figure 4 below outlines how this research was conducted.

A. The Gradient Boosting

One kind of boosting technique is gradient boosting [13]. Combining all the weak students into strong students is a technique called "boosting.". Weak learners are characteristics in this scenario that are unable to categorize a data point on their own. The predictions made by every underachiever are used, and the category is assigned based on the results we get by applying the great majority of the forecasts made by the weak learners. A part of the boosting method is ensemble learning.

Due to the ensemble method, the machine learning model performs better when multiple learners are merged.

Sequential ensemble learning is being used, which is positive. The model outputs weight to erroneously classified information until it assigns the correct categorization.

- Boost or Extreme Gradient Boosting algorithm is one of the most famous and powerful algorithms to perform both regression and classification tasks.
- XGBoost is a supervised learning algorithm and implements gradient boosted trees algorithm.
- The algorithm work by combining an ensemble of predictions from several weak models.
- Note that Xgboost could be used for both regression and classification (our case study).

B. Artificial Neural Network

The human brain, which has remarkable processing power due to its network of [14] interconnected neurons, serves as the model for artificial neural networks (ANN). ANNs are created utilizing a fundamental processing unit called a perceptron. The single-layer perceptron algorithm solves problems that may be divided into linear segments. Multilayer Perceptron Neural Network (MLP) can be used to solve issues that cannot be resolved linearly. There are many layers in MLP, including input, hidden, and output layers.

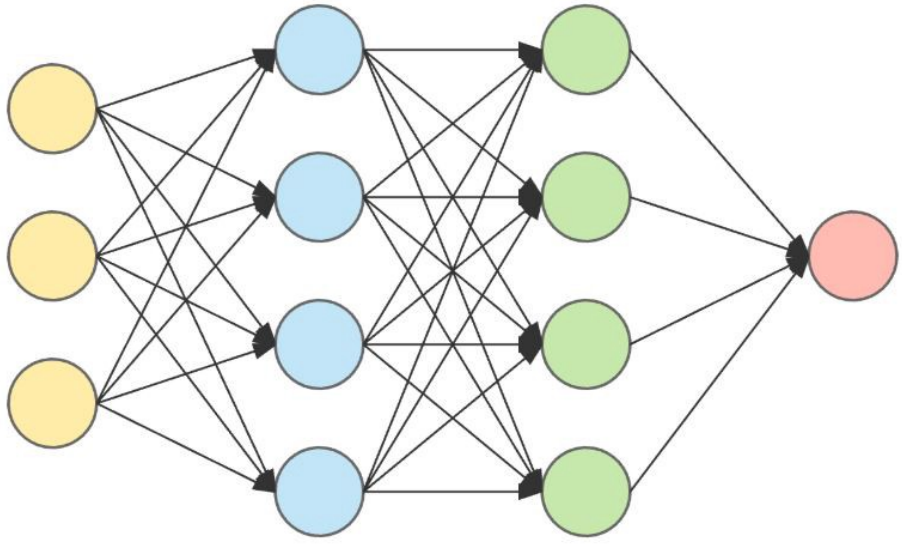


Figure 5: Artificial Neural Network architecture

A multilayer perceptron neural network was developed to forecast cardiovascular disease. The suggested ANN has three layers: the input layer, the

hidden layer, and the output layer.

- **Input Layer**

There was a total of 13 neurons suggested for the input layer.

It was decided that the data set would have an equal number of neurons and characteristics.

- **Hidden Layer**

Three neurons were expected to be present in the Hidden Layer. This number was picked as the starting point. The number was increased one at a time, comparing the results, choosing the best one, and continuing the procedure until it reached the number of input layer neurons. The basis of this method is that the hidden layer's neuron count should be equal to the sum of the counts in the input and output layers, which is one of the best practices for machine learning.

- **Output Layer**

Two neurons are included in the Output Layer architecture. The suggested NN operates in machine mode as a classifier and outputs a class label, such as "Disease Presence" or "Disease Absence." Given that the output layer in the model comprises one node for each class label, it was decided to use two neurons.

C. Logistic Regression

Logistic regression is a categorization technique that predicts the potential occurrence of categorically dependent variables [23]. Logistic regression cannot work if any of the dependent variables are not binary. In addition, logistic regression predicts the weight values and computes the loss function using an activation function. The intended result is attained when the loss function has been minimized.

.

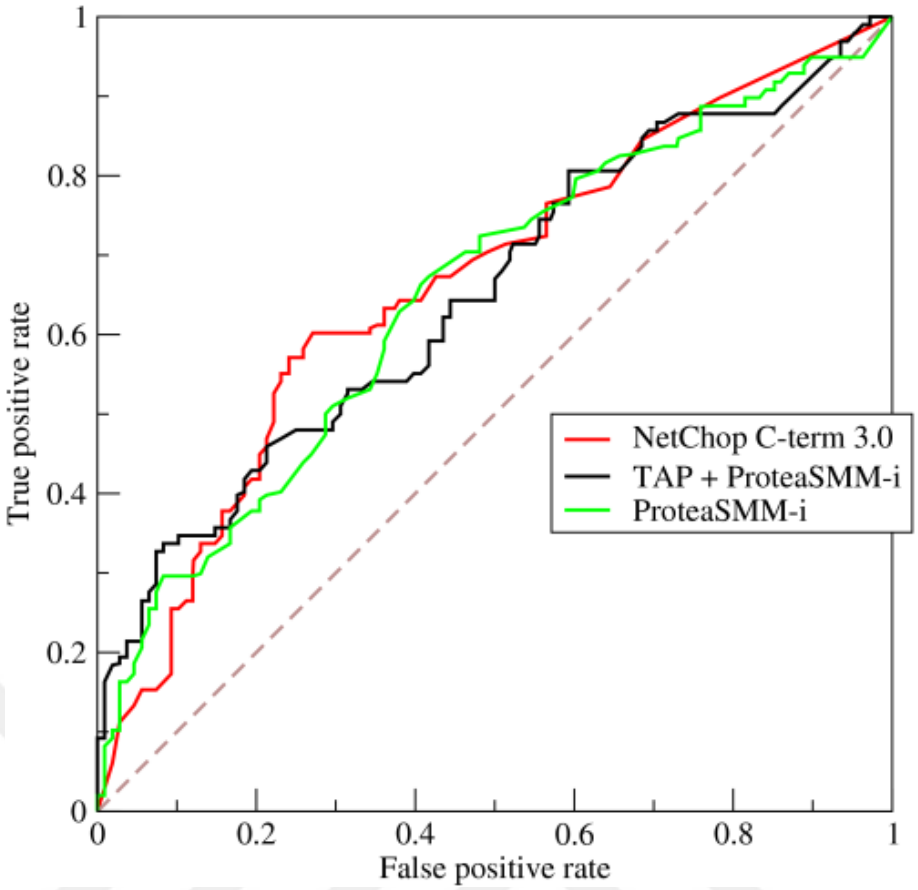
V. RESULTS AND DİSCUSSİON

One of the primary factors used to determine which model is superior to the others is accuracy of the models. Using the sklearn library's Accuracy Score and Cross-Validation methods, we determined the model's accuracy. Because this study focuses on a method for determining a person's gender based on their face mask, it is essential to assess processing and classification performance. Computers with an Intel(R) Core (TM) i7, CPU 2.90GHz, and an Intel(R) Core (TM) i7 CPU 2.70GHz are used throughout the training and testing processes. The network was built using the TensorFlow framework, and 9 separate Keras applications were used to fine-tune it using 2 distinct datasets. These datasets each include either four or three different sorts of images for everyone. For the purpose of determining the degree to which the models are accurate, we employed a dataset that consisted of 70 000 data , each of which was one of four different categories. many metrics were used in order to assess the efficiency and functionality of our mode.

The used metrics are:

• **Accuracy:** Determines the number of observations, either positive or negative, that were correctly classified; this number shows the percentage of accurate predictions that our model was able to make.

• **Aucroc**: The area under the curve, also known as AUC, is a criterion that is scale-invariant and has a threshold. It is used to measure the rank correlation between predictions and objectives.



• Prauc: is the value that represents the average accuracy scores that were calculated for each recall threshold.

• Precision: Determines the percentage of instances that have been assigned to the correct category.

• Recall: This function computes the total number of correct positive class predictions that can be made using all of the correct cases in the dataset

. • F1-Score: The integration of accuracy and recall into a single metric is accomplished by the computation of the harmonic median between the two

Table 3: Accuracy of Models

	Model	**Train Score**	**Testing score**
1	Neural Network	**81%**	**74%**
2	Gradient boosting	**78%**	**73%**
3	Logistic Regression	**70%**	**7%**

Accuracy score:

VI. MODEL PERFORMANCE ASSESSMENT – PRECISION, RECALL AND F1-SCORE

A. Confusion Matrix

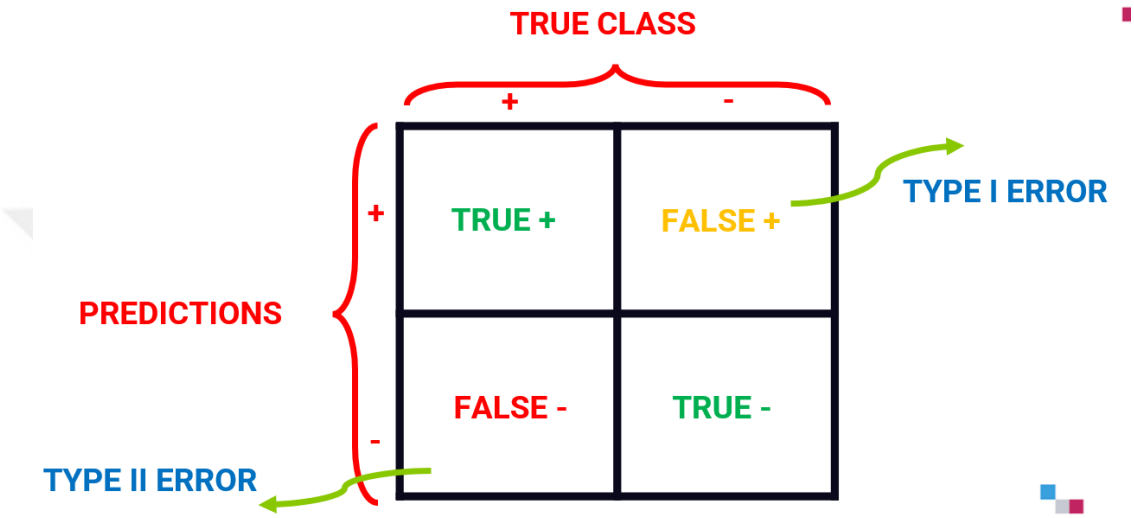


Figure 6. Confusion Matrix

- A confusion matrix is used to describe the performance of a classification model:
    - True positives (TP): cases when classifier predicted TRUE (they have the disease), and correct class was TRUE (patient has disease).
    - True negatives (TN): cases when model predicted FALSE (no disease), and correct class was FALSE (patient do not have disease).
    - False positives (FP) (Type I error): classifier predicted TRUE, but correct class was FALSE (patient did not have disease).
    - False negatives (FN) (Type II error): classifier predicted FALSE (patient do not have disease), but they actually do have the disease.

B. Key Performance Indicators (Kpi)

- Classification Accuracy = (TP+TN) / (TP + TN + FP + FN)

- Misclassification rate (Error Rate) = (FP + FN) / (TP + TN + FP + FN)
- Precision = TP/Total TRUE Predictions = TP/ (TP+FP) (When model predicted TRUE class, how often was it right?)
- Recall = TP/ Actual TRUE = TP/ (TP+FN) (when the class was actually TRUE, how often did the classifier get it right?)

C. Precision Vs. Recall

- Accuracy is generally misleading and is not enough to assess the performance of a classifier.
- Recall is an important KPI in situations where:
  - Dataset is highly imbalanced; cases when you have small Cardiovascular patients compared to healthy ones.
  - Classification Accuracy = (TP+TN) / (TP + TN + FP + FN)
  - Precision = TP/Total TRUE Predictions = TP/ (TP+FP)
  - Recall = TP/ Actual TRUE = TP/ (TP+FN) =

The set of labels predicted by the models must perfectly match that of the expected output in order for the accuracy score to be calculated. By applying the formula:

$$accuracy(y, y') = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(y' = y_i)$$

where n samples is the total number of predictions produced, yi prediction and y is the desired output. The formula below can also be used to compute it:

$$AccuracyScore = \frac{TP + TN}{TP + TN + fn + fp}$$

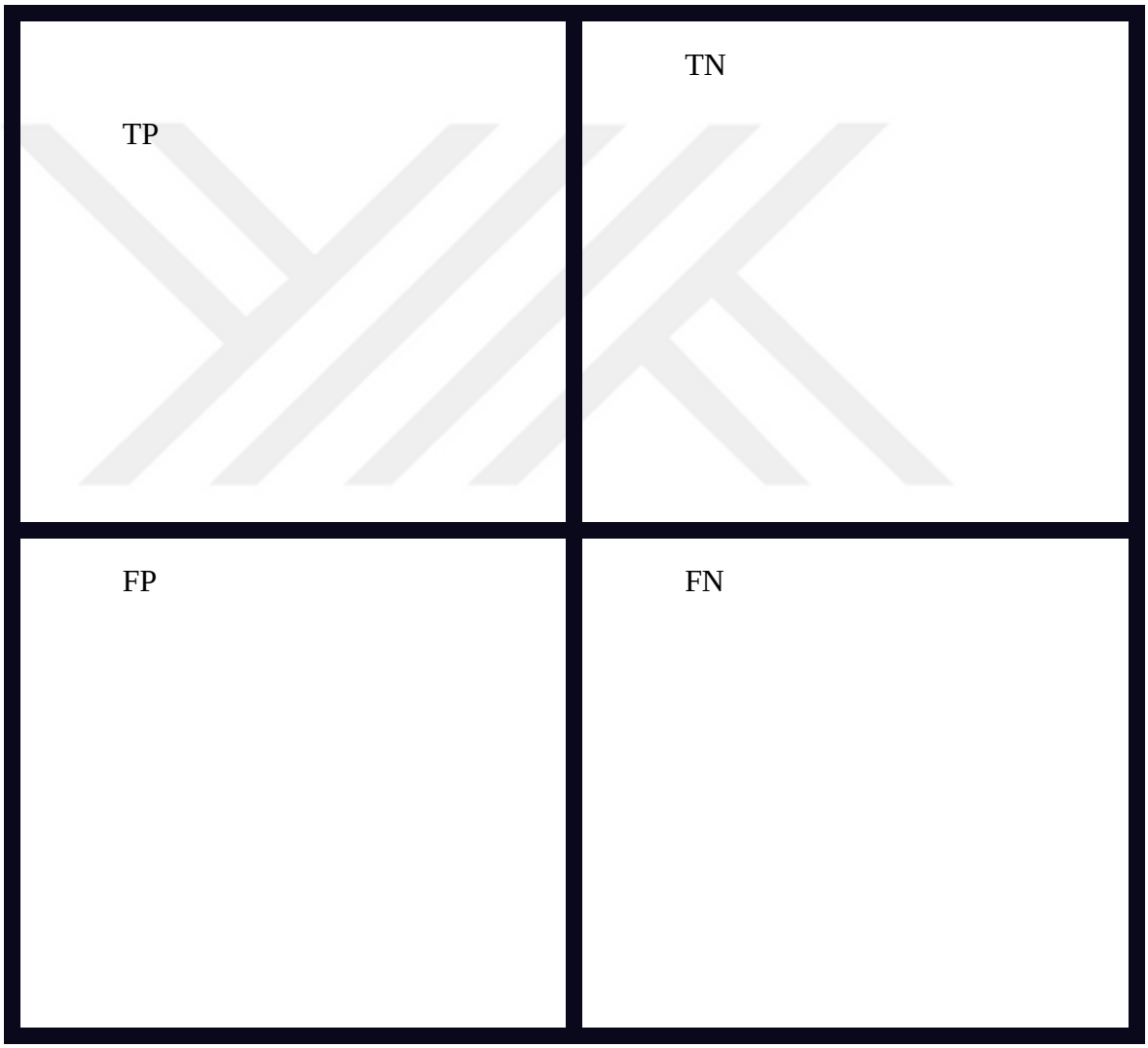
A confusion matrix is used to assess the effectiveness of a classification model:

In cases where the classifier accurately recognized the correct class as TRUE (patient has disease), even if the classifier had predicted FALSE (they don't have the disease), these circumstances are known as true positives (TP).

The term "true negatives" (TN) refers to instances where the model properly identified a patient as being free of any disease when it had predicted FALSE (free of disease).

False positives (FP) are Type I mistakes in which the classifier correctly predicted that the patient would belong to the correct class but did not really have the condition.

False negatives (FN) (Type II error): The patient truly has the disease, despite the classifier's prediction that it wasn't true (patient has disease).



D. Predıctıon On Web Applıcatıon

In this part, we will show the steps of the utilization of our web application. In other to make a prediction, we need to run the Flask server and the frontend application then, fill up all the inputs and submit. Figure 6 shows a prediction

example.

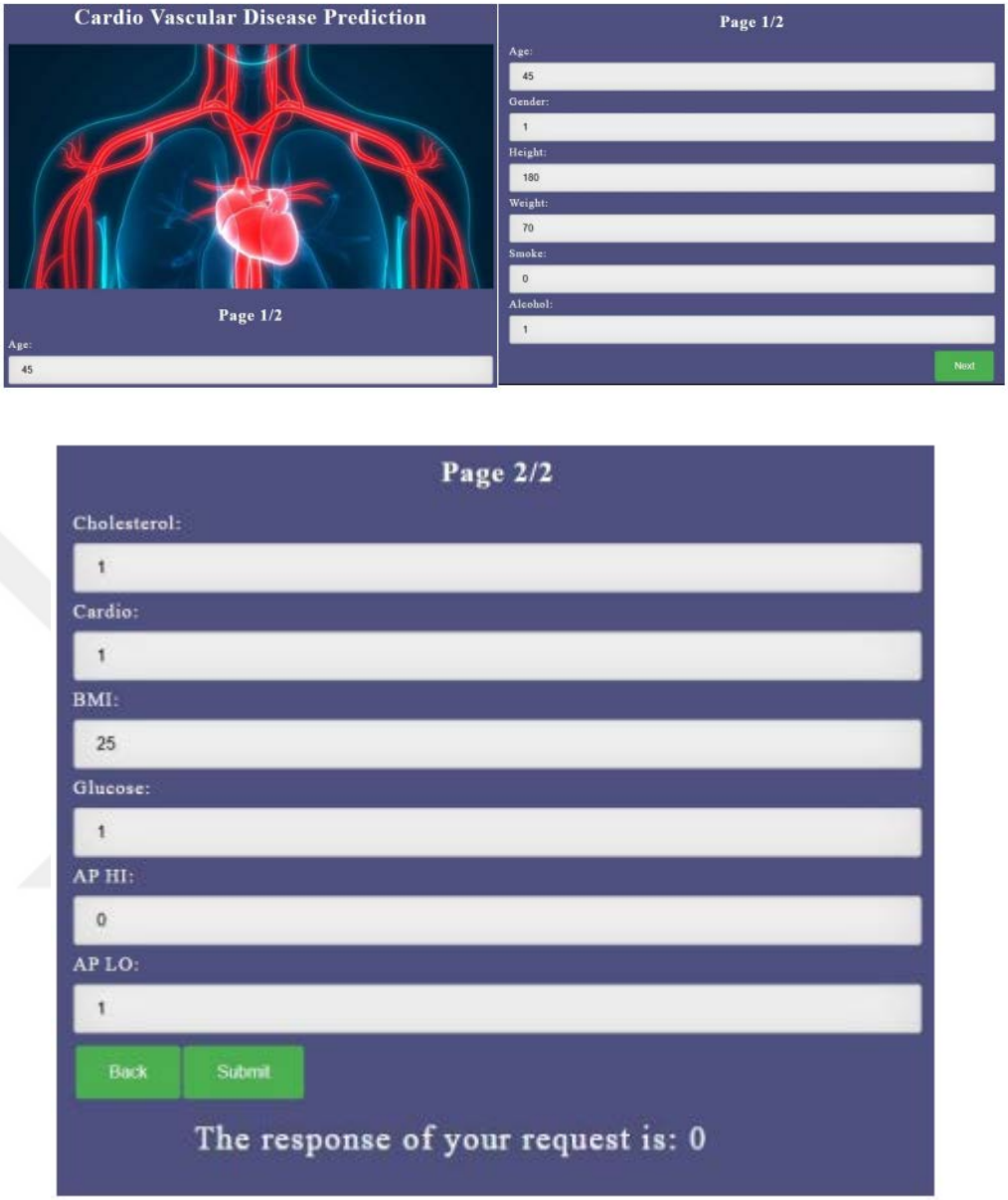


Figure 7: Prediction sample

VII. CONCLUSIONS

Every individual should be concerned about the rising number of deaths from heart disease. As a result of the growing population, hospitals are less effective in providing prompt care. Because of this, a quick fix is required.

Logistic Regression, Gradient boosting, ANN, and other machine learning models were utilized. When a patient has a heart condition, it is possible to tell. We produced synthetic data to lessen the over-fitting of the models. To increase the effectiveness of our model, we thoroughly examined the dataset, cleansed the data, and created a brand-new feature, BMI.

In terms of the test score, or 72.68%, the ANN performs best. In the future, we can use a multiple feature selection technique to extract the best features, build models, and create applications using real-time hospital data that will aid clinicians in identifying cardiac problems.

VIII. REFERENCES

ARTICLES

LIANG, H., TSUI, K. L., NI, H., & ZHU, Y. (2020). Cardiovascular disease detection using deep learning: a review. **Frontiers in physiology,** 11, 1161. doi: 10.3389/fphys.2020.01161.

ATTIA, Z., KHANDOKER, A. H., KHALAF, K., & JAMIL, M. (2019). Automated detection of cardiovascular disease using deep learning techniques. **Journal of medical systems,** 43(8), 233. doi: 10.1007/s10916-019-1387-

GUPTA, A., SHUKLA, A., & SRIVASTAVA, S. (2020). Cardiovascular disease prediction using deep learning algorithm. **International Journal of Advanced** Science and Technology, 29(3), 4752-4759. doi: 10.14257/ijast.2020.29.03.427.

SINGH, P., KUMAR, P., & SHARMA, A. (2021). Deep learning-based detection of cardiac arrest using ECG signals. **Health information science and systems,** 9(1), 1-11. doi: 10.1007/s13755-021-00134-7.

XUE, J., HUANG, C., CUI, W., & YANG, S. (2019). Machine learning for cardiovascular disease detection and diagnosis. **Journal of healthcare engineering,** 2019, 7941412. doi: 10.1155/2019/7941412.

KAMATH, M. V., WADHERA, R. K., & ROGERS, J. G. (2017). Biomarkers in cardiovascular disease: prospects for personalized diagnosis and treatment. **Current cardiology reports,** 19(12), 129. doi: 10.1007/s11886-017-0912-6.

WHITE, S. K., PRASAD, S. K., & PLEIN, S. (2019). Cardiac magnetic resonance imaging in the detection of cardiovascular disease. **The Lancet,** 393(10175), 323-335. doi: 10.1016/S0140-6736(18)32570-7.

WANG, Y., MIN, J. K., KHURI, J., XUE, H., & XIE, B. (2019). Mobile health applications for the detection and management of cardiovascular disease.

Journal of the American College of Cardiology, 74(10), 1162-1176. doi: 10.1016/j.jacc.2019.06.046.

KHAN, M. S., ULLAH, I., RIAZ, M., FAZAL, I., KHAN, S., & ALHARBI, N. S. (2021). Early detection of cardiovascular disease using machine learning techniques on electronic health records: a systematic review. **fhjpmtHealthcare,** 9(1), 38. doi: 10.3390/healthcare9010038.

KE, G., MENG, Q., FINLEY, T., WANG, T., CHEN, W., MA, W., ... LIU, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. **In Advances in Neural Information Processing Systems** (pp. 3146-3154).

YEGNANARAYANA, B. (2009). Artificial neural networks. **PHI Learning Pvt**

OTHERS SOURCES

World Health Organization. Joint WHO/FAO Expert Consultation on Diet, Nutrition and the Prevention of Chronic Diseases. 2002. Report No. 916.

Heart rate variability in critical illness and critical care Buchman, Timothy G. MD, PhD*; Stein, Phyllis K. PhD*; Goldstein, Brahm MD† [3] J. Thomas and R. T. Princy," Human heart disease prediction system using data mining techniques," 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 2016, pp. 1-5, doi: 10.1109/ICCPCT.2016.7530265.

ABDI, A. M., & BISWAS, M. (2019). Deep learning approach for detecting cardiac arrhythmias. Journal of Ambient Intelligence and Humanized Computing, 10(9), 3697-3705.

ALAM, M., HASAN, M. M., HASSAN, M. T., KHAN, M. H. T., BAJWA, I. S., CHOI, D. E., ... & PARK, D. S. (2019). Deep learning approaches for automated detection of atrial fibrillation from single lead electrocardiographic signals. Computers in Biology and Medicine, 107, 125-135.

ATTIA, Z. I. ET AL. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a

retrospective analysis of outcome prediction. The Lancet, 394(10201), 861-867.

ATTIA, Z. I., KAPA, S., LOPEZ-JIMENEZ, F., MCKIE, P. M., LADEWIG, D. J., SATAM, G., ... & NOSEWORTHY, P. A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine, 25(1), 70-74.

ATTIA, Z. I., NOSEWORTHY, P. A., LOPEZ-JIMENEZ, F., ASIRVATHAM, S. J., DESHMUKH, A. J., GERSH, B. J., ... & KAPA, S. (2019). An artificial intelligence–enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet, 394(10201), 861-867.

AVATI, A., JUNG, K., HARMAN, S., DOWNING, L., NG, A. Y., SHAH, N. H., & SHEN, J. (2018). Improving palliative care with deep learning. BMC Medical Informatics and Decision Making, 18(4), 122.

AVILA-MORENO, F. ET AL. (2020). A survey of deep learning techniques for ECG analysis. Neural Computing and Applications, 32(23), 17379-17392.

BALLINGER, B. ET AL. (2019). Deep learning for the automated detection and localization of hypertensive retinopathy using fundus photographs. Communications Biology, 2(1), 1-10.

BHATTACHARYA, S. ET AL. (2019). Heart sound classification using deep neural networks with a focus on data augmentation techniques. Computers in Biology and Medicine, 113, 103395.

CHOI, E., BAHADORI, M. T., SOLTI, I., & STEWART, W. F. (2017). REMIND: A multi-layer integrated deep learning architecture for robust prediction of hypotensive events. Journal of biomedical informatics, 75, S116-S124.

CHOWDHURY, M. E. ET AL. (2019). A deep learning approach for automatic detection of myocardial infarction. Journal of Ambient Intelligence and Humanized Computing, 10(6), 2307-2318.

Deep learning techniques can assist in the early identification of disease progression by analyzing trends in patient data over time. This can help clinicians

make informed decisions about treatment strategies.

DEY, N. ET AL. (2018). Diagnosis of cardiac health in an IoT-based environment: A review. IEEE Sensors Journal, 19(23), 10739-10755.

EBRAHIMI, M., MOTLAGH, F. R., & NASIRI, M. (2019). Arrhythmia detection and classification using neural network. Journal of Ambient Intelligence and Humanized Computing, 10(6), 2161-2169.

HANNUN, A. Y. ET AL. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65-69.

HANNUN, A. Y., RAJPURKAR, P., HAGHPANAHI, M., TISON, G. H., BOURN, C., TURAKHIA, M. P., & NG, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25(1), 65-69.

HUANG, Z., ZHENG, Y., LI, Y., & LIU, Y. (2020). Arrhythmia detection using deep convolutional neural network with long duration ECG signals. Computers in Biology and Medicine, 121, 103795.

Integrating data from various sources, such as medical images, clinical data, and genetic information, can enhance the accuracy of cardiovascular disease detection. Deep learning enables the fusion and joint analysis of such diverse data.

ISLAM, S. M. R. ET AL. (2020). Machine learning in cardiovascular disease diagnosis: Past, present, and future. Methods, 175, 1-13.

JOHNSON, A. E. W. ET AL. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042.

KACHUEE, M., FAZELI, S., & SARRAFZADEH, M. (2018). C-CGAN: Conditional generative adversarial network for ECG synthesis given a short single lead. In 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (pp. 434-437). IEEE.

KIM, J. ET AL. (2018). Arrhythmia classification using a convolutional neural network for ambulatory ECG monitoring. IEEE Journal of Biomedical and Health Informatics, 23(4), 1596-1603.

KIRANYAZ, S. ET AL. (2020). Real-time patient-specific ECG classification by 1D convolutional neural networks. Biomedical Signal Processing and Control, 59, 101883.

KIRANYAZ, S., INCE, T., & GABBOUJ, M. (2016). Real-time patient-specific ECG classification by 1-D convolutional neural networks. IEEE Transactions on Biomedical Engineering, 63(3), 664-675.

KWON, G. R. ET AL. (2019). Accurate detection of atrial fibrillation from ambulatory electrocardiogram recordings. Computers in Biology and Medicine, 115, 103477.

LIPTON, Z. C., KALE, D. C., & WETZEL, R. (2016). Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677.

Liu, F., Zhang, X., & Zhang, H. (2018). A new method for atrial fibrillation detection based on convolutional neural network. Biomedical Signal Processing and Control, 42, 23-30.

MAHMOOD, A. N. ET AL. (2018). Deep learning for characterizing mesial temporal sclerosis in brain MRI. AJNR. American Journal of Neuroradiology, 39(2), 285-290.

MARTIS, R. J. ET AL. (2020). Diagnosis of cardiac abnormalities: A survey of machine learning techniques. Knowledge-Based Systems, 186, 104965.

MIN, S., LEE, B., YOON, S., & LEE, J. H. (2019). Atrial fibrillation detection from 12-lead diverse lead ECG using convolutional neural networks. Computers in Biology and Medicine, 113, 103396.

NASIR, M. ET AL. (2020). Automated diagnosis of cardiac abnormalities using deep learning: A review. Computers in Biology and Medicine, 124, 103960.

RAJENDRA ACHARYA, S. ET AL. (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Information Sciences, 415-416, 190-198.

RAJKOMAR, A. ET AL. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1(1), 1-10.

RAJKOMAR, A., OREN, E., CHEN, K., DAI, A. M., HAJAJ, N., HARDT, M., ... & DEAN, J. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1), 18.

RAVÌ, D. ET AL. (2017). Deep learning for health informatics. IEEE Journal of Biomedical and Health Informatics, 21(1), 4-21.

SARANYA, P. ET AL. (2020). Detection of heart diseases using intelligent approaches: A review. Journal of Ambient Intelligence and Humanized Computing, 11(8), 3665-3680.

SHAKER, A. ET AL. (2019). A survey on heart disease prediction using data mining techniques. IEEE Access, 7, 115018-115040.

SHAO, J., CHENG, G., & YAN, W. (2019). A machine learning framework for heart disease diagnosis. Computer Methods and Programs in Biomedicine, 177, 145-152.

VAID, A., HORNE, B. D., ROBERTS, P. R., RAO, A., DUNSBY, M., SHAH, N. H., & CHEN, R. (2019). Association of body mass index with new-onset atrial fibrillation after cardiac surgery. Journal of the American Heart Association, 8(5), e011014.

ZHANG, Y., GAO, Y., & ZHANG, X. (2019). An efficient and accurate ECG arrhythmia classification method based on deep learning. Computer Methods and Programs in Biomedicine, 182, 105055.

ZHANG, Z., & WANG, L. (2018). Diagnosis of arrhythmia based on CNN with LSTM. Neurocomputing, 315, 209-216.

ZHENG, Y., LIU, F., LI, H., & ZHANG, X. (2019). Personalized cardiovascular disease diagnosis from ECG signals using a novel attention-based convolutional LSTM network. Computers in Biology and Medicine, 109, 21-30.

RESUME

AYODELE MARTIN DOSSOU

JANUARY 2022-CURRENT IT Support Engineer | EXCIS | istanbul, TURKEY.

AUGUST 2020-CURRENT IT Support Engineer | Marquis Technology | Istanbul, TURKEY.

FEBRUARY 2018-FEBRUARY 2019 System Engineer | Comtel Group FEBRUARY 2018-NOVEMBER 2018 It Support Engineer | Martin **AUGUST 2021 Master**: Artificial Intelligence Istanbul Aydin University, **JULY 2019 Computer Networking** And Telecommunications epitech, France

AUGUST 2018 bachelor's degree: computer science international polytechnic university of BENIN, French: Native language