Bachelor of Science in Computer Science & Engineering

Hypertension Risk Prediction and Preliminary Medical Suggestion Using Machine Learning

by

Md. Zubair

ID: 1504063

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

April, 2021

Hypertension Risk Prediction and Preliminary Medical Suggestion Using Machine Learning

Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Md. Zubair
ID: 1504063

Supervised by

Dr. Muhammad Ibrahim Khan
Professor
Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)
Chattogram-4349, Bangladesh.

The thesis titled '**Hypertension Risk Prediction and Preliminary Medical Suggestion Using Machine Learning'** submitted by ID: 1504063, Session 2019-2020 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

Chairman

Dr. Muhammad Ibrahim Khan
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (Ex-Officio)

Dr. Asaduzzaman
Professor & Head
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (External)

Dr. Kaushik Deb
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Signature of the candidate

Date:

Acknowledgement

It is very difficult to complete a task without the mercy of almighty and the input, support and guidance of many individuals and institutions. At the very beginning I would like to extend my limitless praise, obligations and gratitude to almighty, the entirely merciful, gracious and beneficent, whose propitiousness enabled me to complete this thesis successfully. I would like to express my special thanks of gratitude to my teacher, supervisor Dr. Muhammad Ibrahim Khan, Professor, Department of Computer Science and Engineering, Chittagong University of Engineering and Technology (CUET) for encouraging me and also for his continuous guidance, cooperation, constructive criticism and helpful suggestion in carrying out the thesis.Without his constant and intense cooperation I could not complete this thesis within the time-frame. I acknowledge with thanks the kind of patronage, loving inspiration which I have received from my supervisor.

I owe my gratitude to Dr. Syed Ahmed Abdullah Lecturer, Oral Microbiology Shaheed Suhrawardy Medical College who helped me a lot in figuring out the problem in medical domain. He also provided all kind of necessary information regarding the thesis.

Finally, my appreciation and heartfelt gratitude are extended to each and every person who somehow helped us in this regard.

Abstract

Hypertension is one of the global health issues. People around the globe are largely suffering from cardiovascular diseases among them a significant amount of people are suffering from hypertension.If it is not taken into account at it's early ages, may lead to other severe cardiovascular diseases. At the early stages of hypertension, people doesn't care about the consequences of hypertension. Most often the mass people doesn't conduct any checkup for hypertension. So, a system which can predict the risk of hypertension from non-clinical data and providing preliminary suggestions based on the user data might help the people. A machine learning based system has been developed for predicting the hypertension and preliminary suggestion has been given accordingly. The model has been compared with several machine learning algorithms like artificial neural network, random forest, logistic regression, naive bayes, decision tree, k-nearest neighbour, SVM. The random forest model works best for the hypertension data.

Keywords: Hypertension, cardiovascular, disease, non-clinical data, machine learning, random forest, preliminary suggestion,

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Blood pressure is the force imposed by circulating blood against the walls of the body's arteries, the major blood vessels in the body. Hypertension occurs when blood pressure is too high [1]. Hypertension is also known as high blood pressure.

Hypertension is one of the global public health issues. Each year about 17 million deaths are caused by cardiovascular diseases. Among them hypertension is the cause of 9.4 million deaths [2]. At least 45% deaths due to heart disease and 51% deaths due to stroke are related to hypertension [2]. Hypertension causes 1 out of every 8 deaths and considered as the third leading killer in the world [3]. The projected growth of hypertension from 2000 to 2025 is 26% to 29.2% [4]. Research shows that the developing countries' condition is more vulnerable. 20% of the adult and 40% - 65% elderly people of the developing country are suffering from hypertension. One third of the hypertension patients do not check their blood pressure in their whole life [5]. Because mass people cannot recognize the effect of hypertension in the early stage of it [2].

To mitigate the risk and helping the mass people, a system has been designed to predict the risk of hypertension along with preliminary medical suggestion based on the user data. The upcoming sections will provide a brief overview of the model and other directions of the works.

1.1.1 Hypertension and Predictive model

Hypertension is one of the cardiovascular diseases. It depends highly on personal lifestyle and sometimes it is genetical. As the disease is highly dependent on the non clinical variables, it can be predicted easily by intelligent system like machine learning.The model can provide the solution only by entering the non clinical data. More precisely, the system would be a great assistant for the human being.

There are some risk factors which mainly accountable to the hypertension. The parameters are many. Not all of them are equally responsible to hypertension. Feature extraction is highly needed in that case to find out the important features.

1.2 Design Overview of the Hypertension Risk Prediction Model

For any prediction mode, the most crucial part is meaning full and reasonable data. Without data, a predictive model doesn't make any sense. For the model, I have tried to collect real-time data by myself. I had collected some of the data. Unfortunately, for the sudden outbreak of COVID-19, I have failed to collect enough data needed to train the model.So, I have ended up training the model with built in dataset. Some data prepossessing is required to fit the model.

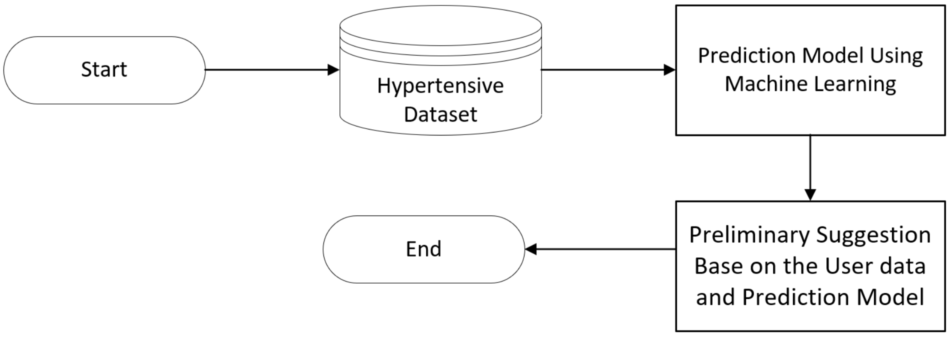
Start
Hypertensive Dataset
Prediction Model Using Machine Learning
Preliminary Suggestion Base on the User data and Prediction Model
End

Figure 1.1: Overview of the proposed model.

The overall process is shown in details with the figure 1.1. In the prediction model, several machine learning techniques are compared. And found the random forest classifier is the best for the model. Finally, the system provides some preliminary suggestions based on the user data and the prediction model.

1.3 Difficulties

Creating a useful predictive machine learning model is always challenging. First challenge is finding out the proper dataset which can be used to train a machine learning model. Finding out the best model is also a difficult one. And finally most difficult one is providing some suggestions based on the user behaviour.

1. **Finding out the proper dataset**: There are some available data of hypertension. But the main problem is that the the dataset are not that much useful in case of real-time prediction model. So, I have tried to create the dataset. Unfortunately, it has hindered the global pandemic of COVID-19. The questionnaires have been selected with the help of expert doctors.

2. **Training the best Machine Learning model**: As the hypertension can be classified into different stages, we can predict the stages of the hypertension using classification algorithms of machine learning. For the sake of finding out the best model, I have trained the model with the popular machine learning classification algorithms. And the best model has been selected with comparing the evaluation matrices.

3. **Suggestion generation based on prediction and user data**: Generating suggestions automatically is a difficult issue. For doing this, we need to have proper guidelines of the medical experts. I along with some of the medical experts figure out the possible suggestions. As it is a machine generated output, no medicine has been suggested for safety issue.

1.4 Applications

The model has been created to evaluate the potential risks of hypertension and some preliminary suggestions from non-clinical dataset. So, I personally believe the proposed model can be used for multiple dimensions. Some of the possible applications are mentioned below–

- General people for their hypertension risk check.

- For getting preliminary medical suggestions for hypertensive patients without any clinical data.
- The model can also be included in the intelligent medical assistance websites.

1.5 Motivation

Behind the creation of the model, there is a noble issue. In each and every year, a huge number of people are suffering from hypertension. Some statistics have shown in the 1.1 section. At the early, stages of hypertension we don't feel the consequences of the disease. In developing countries like ours', people are not that much conscious to check their body condition. But Hypertension may lead to various severe diseases like heart attack or stroke, aneurysm, heart failure, weakened and narrow blood vessels in kidney, thickened, narrowed or torn blood vessels in the eyes, metabolic syndrome, trouble with memory or understanding, dementia and so one [6].

If people use this system to evaluate the hypertensive condition, they will have a overview of their health condition and preliminary suggestions. It will help the people to be cautious about their health. We should keep in mind, "prevention is better then cure." Hopefully, this model will be a great assistance to lessen the risk of hypertension.

1.6 Contribution of the thesis

Every thesis or research has been conducted to find out a unique contribution. It is not some thing like integration or gathering some information rather it is something like innovation, pointing out some cause and effect or proposing a model which would be beneficial for human being. The main contribution of the thesis works are as follows:

1. Finding out the important variables for hypertension.

2. Choosing the best predictive machine learning model for the hypertension dataset.
3. Automatically generate the preliminary medical suggestion based on user data and machine learning prediction model.

1.7 Thesis Organization

Organization of the rest of the thesis report is as follows -

- **Chapter - 2** provides the a brief summary of the previous works in the related fields and it's background.
- **Chapter - 3** represents a detail information and general procedure of the proposed model.
- **Chapter - 4** is all about the evaluation and experimental outcomes of the proposed system.
- **Chapter - 5** concludes the thesis findings. Additionally a brief direction of future work has been given.

1.8 Conclusion

In this chapter, basic introduction of our work has been introduced. Carried out a task like this is not that much easy. The basic overview of the hypertension risk prediction model has been explained in detail. Motivation behind the thesis, difficulties I have faced have been narrated as well. The contribution of our work has also been discussed. In the next section, the current state and background of the problem have been stated.

Chapter 2

Literature Review

2.1 Introduction

Hypertension also known as high blood pressure to the common people. It is not something new rather it is one of the most common health issues around the globe. Field of medical science is one of the promising fields to apply technology. Technology is the assistant of human being. If it can be applied for the betterment of the human, it will be a huge contribution.

Scientists and researchers around the world are trying to apply new technologies for the automation of medical sector. Among the many technologies machine learning is one of the most effective system to predict the risk of any disease. Over the years many individual researches have been conducted upon the extracting the risk factor of the hypertension.

2.2 Related Literature Review

Most of the cases some statistical approaches are used to extract the risk factor. Hypertension and cardiovascular risk factor are extracted from the patient data [7]. The research is done on the conventional statistical method.It doesn't provide a meaningful insights into the hypertension risk prediction.Kearney at el. [8] showed impact of hypertension on global data.The main target of the work was to show a direction to prevent hypertension.There was no aim to create any model for prediction of the hypertension. In [4] the analysis of anti-hypertensive drug suggestion has been done. No predicting model has been suggested for the anti-hypertensive drug.

S. Mohan at el. [9] proposed a machine learning model for predicting heart diseases. The research finding showed an improved method of selecting features before applying the machine learning model on the dataset. It had improved the accuracy. Prediction of diabetes and cardiovascular disease with clinical data have been shown by the researchers. They also showed the comparative analysis of machine learning algorithms for predicting the diabetes and cardiovascular diseases. Moreover, no suggestion had been given based on the prediction model [10].

Some researches have been done in the recent years for predicting hypertension risk of individuals using machine learning.

Evaluation and comparison of machine learning techniques for hypertension predicting the individuals at risk of developing hypertension [11]. It only shows which model is the best to the apply machine learning. Some important factors are missed out in the dataset. An article shows the prediction of increased blood pressure using machine learning [12]. Some parameter like BMI, waist circumference, hip circumference, waist-hip ratio etc. are considered. Using Canadian4 Primary Care Sentinel Surveillance Network (CPCSSN) dataset, predict hypertension with machine learning algorithms [13]. It takes a very few variables for the prediction. Hypertension prediction with radiology image has also done to detect pulmonary hypertension [14].

2.3 Conclusion

Hypertension is one of the common health issue over the world. It leads to many other cardiovascular diseases. Over the years many researchers have been worked to analyze the hypertensive data, and other heart disease data. Many predictive model have been introduced. But my research findings has shown the best model for predicting the hypertension risk as well as preliminary medical suggestions. Preliminary medical suggestion is our unique contribution. In the upcoming sections, the research methodology and other comparative analysis have been introduced.

Chapter 3

Methodology

3.1 Introduction

Methodology is the best of the bread of any research or scientific writings. It simply provides answers of the question "how" of any research. Particularly, it's all about how a researcher systematically approaches to ensure the acceptability of the work through valid and reliable results.

For the proposed model. I have gone through several process. First of all, data have been collected. A little bit research has done for the making questionnaires. Though the collection of data has not been fulfil for the world wide pandemic situation of COVID-19. At last, I was bound to work with healthcare heart study dataset [15]. And then machine learning model creation, evaluation and preliminary medical suggestion generation.

3.1.1 Questionnaires Development

The parameters are many. Not all of them are equally responsible to hypertension. Some main variables/ risk factors are chosen for the preceding thesis. The risk factors are gender, family history, demographic area, occupation, food habit, marital status, mental health, drug addiction, tobacco use, smoking history, waist circumference, BMI, education, physical activity, disease condition, systolic blood pressure, diastolic blood pressure [2, 3, 5, 6, 7, 12, 16]. With the help of the research articles and consulting with medical expert a questionnaires has been developed. Dr. Syed Ahmed Abdullah[1] is one of the expert who helped to figure out questionnaires. The questions are as follows-

[1] Lecturer, Oral Microbiology Shaheed Suhrawardy Medical College

1. Age
2. Gender (i.Male ii.Female iii.Third gender)
3. Demographic area of living (i.Urban ii.Rural)
4. Occupation (i.Employed ii.Student iii.Unemployed iv.Retired v.Housewife vi.Businessman)
5. Education (i.No education ii.Primary iii.Secondary iv.College or Higher)
6. Family History (i.HTN present ii.Not Present)
7. Marital status (i.Unmarried ii.Married iii.Divorced/Widow/Widower)
8. Mental Health (i.Normal ii.Anxiety iii.Depression iv.Nervous v.Mental Pressure)
9. Food habit (i.Junk food ii.Normal)
10. Smoking History (i.Currently Smoking ii.Previously Smoking iii.Never Smoke)
11. Substance Abuse (i.Addicted ii.Previously addicted iii.None)
12. Tobacco Use (other than smoking) (i.In use ii.Previously Smoking)
13. Waist circumference Ratio
14. 15. Disease Condition (i.Diabetes ii.IHD / Coronary Disease iii.Heart Failure iv.Asthma / COPD (emphysema, chronic bronchitis, and refractory asthma) v.Kidney disease (AGN, CKD) vi.Cancer vii.Stroke/CVD viii.Dyslipidaemia (↑ Total Cholesterol, ↑ TG = Triglyceride, ↑ LDL, ↓ HDL))
15. BMI
16. Systolic blood pressure
17. Diastolic blood pressure

These information are used to create dataset.

3.1.2 Sample Data

Some data has been collected from a field survey. For collecting the data, digital blood pressure measurement machine is used. BMI measurement has be taken with weight machine and measurement tape. And other data are taken with the direct interview of the patients.

Table 3.1: Sample data for hypertensive patients

Age	Gender	Demographic Area	Occupation	Educational Qualification	Food Habit	BMI	Hypertension Type
56	Male	Urban	Employed	4	Junk	25	Prehypertension
54	Female	Urban	Retired	4	Normal	31	Mild
23	Female	Rural	Student	4	Normal	21.6	Normal

Table 3.1 shows the sample data with a few attributes. Because all attributes can not be accommodated within the page margin.

3.2 Steps Associated with the Proposed Method

The proposed method consists of two parts. First one is the prediction of the hypertension and the second one is providing suggestions according to the user data and the result from the prediction model.

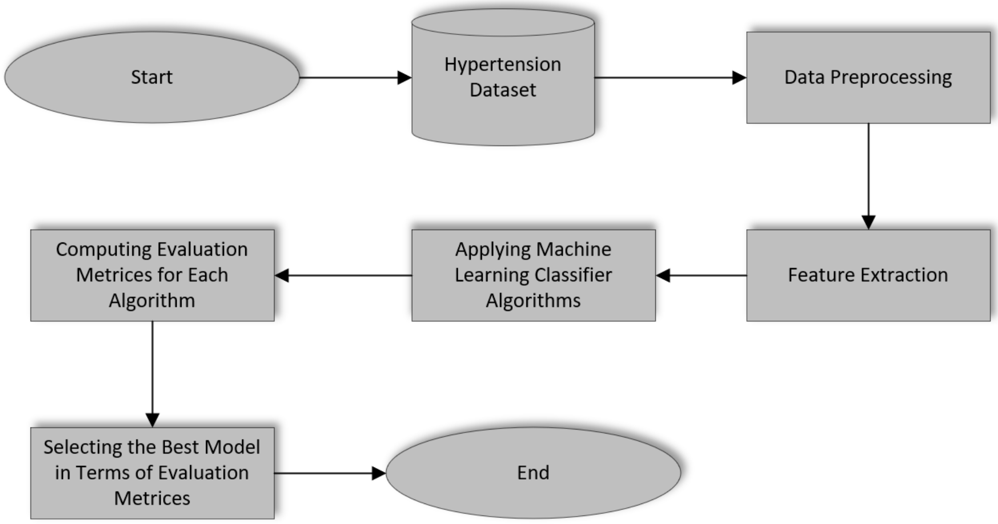
Start
Hypertension Dataset
Data Preprocessing
Feature Extraction
Applying Machine Learning Classifier Algorithms
Computing Evaluation Metrices for Each Algorithm
Selecting the Best Model in Terms of Evaluation Metrices
End

Figure 3.1: Prediction steps for the proposed method.

The steps shown in figure 3.1 refers the prediction steps for the hypertension risk prediction.

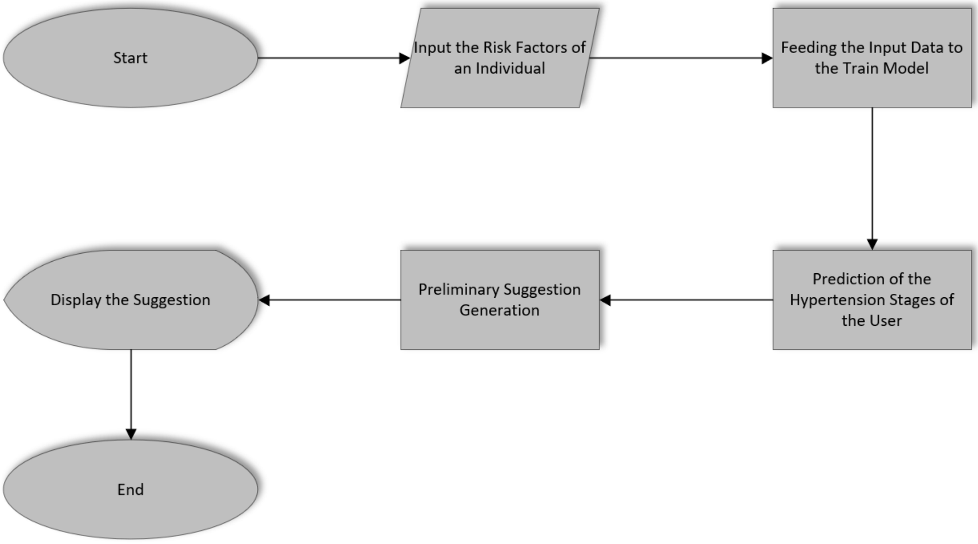
Start
Input the Risk Factors of an Individual
Feeding the Input Data to the Train Model
Prediction of the Hypertension Stages of the User
Preliminary Suggestion Generation
Display the Suggestion
End

Figure 3.2: Preliminary suggestion steps for the proposed model

The steps shown in figure 3.2 indicates the steps associated with the suggestion part. In this part, the system takes input form the user and feed the input data to the previously trained model shown in 3.1. The prediction model provides a prediction of the user hypertension risk. With the help of that result and other inserted data the proposed model automatically generates the suggestion for the user.

3.3 Explanation of the Proposed Method

In this section, the explanation of the proposed method has been given. Behind the main implementation there are so many techniques and algorithms have been implemented. All of the basics will be discussed in this section.

3.3.1 Hypertensive Data Prepossessing

Machine learning model is a mathematical model. These models are the combination of mathematical and statistical concepts. We know that both statistical and mathematical models work on the numerical data. So, one of our main target is to convert the non numeric data to numeric data. As our predictive model is one of the categorical problem. Conversion of continuous data to discrete categorical data helps the classification algorithms to perform better. Some basic data prepossessing strategies which have been implemented in the proposed methods are given below-

1. Converting the non numeric data to numeric data.
2. Categorize the continuous numeric data to discrete categorical data.
3. Fill the null values by choosing the best filing techniques.
4. Omit the rows and columns with a large number of null values.
5. Normalization of the attributes.

Another important prepossessing is categorizing the stages of hypertension from the systolic and diastolic blood pressure. The blood pressure may be categorized into different stages [16] as follows.

- STAGE 1 or Prehypertension is 120/80 (mm.Hg) to 139/89 (mm.Hg)
- STAGE 2 or Mild Hypertension is 140/90 (mm.Hg) to 159/99 (mm.Hg)
- STAGE 3 or Moderate Hypertension is 160/100 (mm.Hg) to 179/109(mm.Hg)
- STAGE 4 or Severe Hypertension is 180/110 (mm.Hg) or higher

3.3.2 Feature Extraction

There are so many variables which have an impact on the hypertension. All of the variables are not equally important. Though we have done some preliminary researches for selecting the features, we need to do further research for finding out the the important features with systematic analysis [17].

$$\mathbf{R}_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{3.1}$$

Where,

R_{xy} represents the correlation between two features x and y.
x_i values of $x-$ variable in a sample.
$\bar{x}$ indicates the mean value of x variable.

y_i values of $y-$ variable in a sample.
$\bar{y}$ indicates the mean value of y variable.

I have considered the values based on the correlation. The higher the correlation value the higher it's importance. The features which holds lower correlation value are discarded.

3.3.3 Machine Learning Algorithm

Machine learning is a process of learning and adopting with the help of training data without explicit instruction. There are basically two types of machine learning techniques

1. Supervised
2. Unsupervised

Supervised Learning is the technique of learning form a level data where for a input data the output is given. And the algorithms work accordingly.

Unsupervised Learning is the technique of learning where the operation or training is done on the unlabeled dataset. The algorithm groups the similar types of data for finding out the similar features.

Anyway, our model is created based on the labeled data. So, the problem can be formulated as a supervised learning problem. The supervised learning algorithms are of two type.

1. Classification
2. regression

Classification algorithms work on the categorical output values where each output can be assigned in a class. The algorithms works for discrete output values.

Regression algorithms work on continuous regressive output values. Here, the output is continuous value.

Our model predicts the categorical output. So, it can be said that it is a classification problem. In the upcoming subsections, different types of machine learning classification algorithms will be discussed.

3.3.3.1 Logistic Regression

It is one of the most commonly used classification algorithms. Logistic regression is a binary classification algorithm. Basically, it works with a cost function and gradient descent algorithm helps to minimize the cost[18]. Finally, classification is done with a sigmoid function. The output value of the sigmoid function fluctuates between 0 to 1. If the output value of the sigmoid function is greater than or equal to threshold value (like 0.5) the class of the instance belong to the positive class otherwise it is negative class.

3.3.3.2 K-Nearest Neighbour

K-nearest neighbour (KNN) classifier is a simple and widely used classification machine learning algorithm. It performs well when the dataset size is small.

When the algorithm goes for operation, it calculates all the distances from a given data which will be classified. Then it calculates distances of other training from the data given data. And selects only K numbers of training data.Finally, it assign the class of the data from the majority class of k training data [19].It has no training phase. That's why it is sometimes known as lazy algorithm.

3.3.3.3 Naive Bayes

Naive Bayes classifier is a simple classifier. It is a probabilistic classifier. Mainly it works on Bayes theorem which refers strong independence assumption into the features. Using Bayes theorem, the classifier calculates the probability for negative and positive outcomes. Finally, the probability for negative and positive outcomes are compared. If the positive outcome is greater than negative outcomes then the class is positive and vise versa [20].

3.3.3.4 Support Vector Machine

Support Vector Machine (SVM) classifier is a strong machine learning classification algorithm. By nature it is a binary classifier but multiple classification can be done by it. It assumes a hyperplane which divides two classes. The hyperplane is selected in a systematic way. Two margin is assumed at the edge of two classes and finally a line is drawn at the very middle of the two margins. And the line is known as hyperplane. This hyperplane helps the algorithm to classify [21].

3.3.3.5 Decision Tree

Decision tree is another classification machine learning algorithm algorithm. It works well on categorical data. It is a tree like structure. There is leaf and root nodes. The leaf and nodes are selected with the help of information gain. Nodes are used to make decision and the leaves provide the final output or class [22]. It also works for missing values as well. It is effective classification algorithm.

3.3.3.6 Random Forest

Random forest classifier is supervised machine learning classification algorithm. The word "forest" means it is consisted of so many trees. More precisely, it is the combination of decision trees. It is a ensemble method of decision tree. Mainly, combination of trees increases the overall accuracy of the model. The main concept is building multiple decision trees and merging together for having more accurate result [23].

3.3.3.7 Artificial Neural Network

Artificial Neural Network (ANN) is the most sophisticated machine learning algorithm for classification and regression model. It is a layered network like structure. There are three types of layers input layer, hidden layer and output layer. Input values are passed through the input layer and hidden layer works to select features and other calculations with weighted values. In each network there is an activation function . And the final output layer assign a class to the input value based on the input data [24].

3.3.4 Implementation of the Proposed Method

It is the final step combining all the concepts and models. With the accumulation of the other process, this section will provide a sequential guideline. The whole implementation process can be divided into two parts. One part is creating a prediction model and the second one is providing suggestion based on the prediction model and the user data.

3.3.4.1 Steps for Prediction Model

In this subjection, the the process of creating hypertension risk prediction model will be discussed in brief. Figure 3.1 depicts the overall steps of prediction model.

- **Step -1:** In this step the input *Hypertension dataset* has been collected. Very beginning I have tried to collect the dataset. Unfortunately, for coronavirus pandemic I have failed to collect sufficient data for training the model. Later on the healthcare heart study dataset [15] is used for training the model.

- **Step -2:** Machine learning models are based on statics and mathematics. So, it can only be operated on the numerical data. It is worth mentioning that the dataset is not fully numerical. So, the dataset is needed to be convert to numerical values. And missing values are filled accordingly.

- **Step -3:** Feature extraction is one of the most important part of a machine learning model. There are many features which is related to hypertension. But all of them are not equally important for the predictive model. So, the important features has been selected by calculating the correlation among the features. Finally, the top correlated features have been selected for traning the model.

- **Step -4:** In this step, the model is trained with available classification machine learning models. The model is trained with logistic regression (LR), k-nearest neighbour (KNN), suport vector machine (SVM), decision tree (DT), random forest (RF) and artificial neural network (ANN). All of these machine learning algorithms are used to train the model and accuracy is compared.

- **Step -5:** Then the evaluation matrices have been measured for evaluation of the models. The main target is to find out the best classifier for our hypertension dataset.
- **Step -6:** It is the final step of model selection. Based on the previous evaluation matrices the best model has been selected. Now, with the help of grid search method we have tuned the parameters using grid search [25] for having the best result out of the classification algorithm.

3.3.4.2 Steps for Preliminary Suggestion Model

Providing suggestion based on the user data is quite difficult and there might be many risks. It is worth mentioning that no medicine is suggested here for human. A dictionary has been created for preliminary medical suggestion. For information, Guyton and Hall textbook of medical physiology is used [26].

- **Step -1:** In this step, the system ask to insert the risk variables of the user. The variables are as same as the training variables.These variables are used to feed in the prediction model.
- **Step -2:** From the input data of step -1 the system saves the values. These values are used to feed into the trained machine learning classification model. And the model generates a predictive stage of hypertension.
- **Step -3:** Now the system receives the predictive stage of hypertension from the step-2 and other risk factors from step-1 of the user. Finally, the proposed model will matches the inputs and hypertension stages to the dictionary. And generates suggestion automatically.

3.4 Conclusion

In this chapter, detailed explanation of the proposed method has been given. In a nutshell, the proposed method is divided into two part one for prediction of the hypertension risk and another is the preliminary medical suggestion. The best classification model is used to predict the stage of hypertension. For preliminary medical suggestion medical field experts suggestion have been considered. In the upcoming chapter, experimental results and other and some discussion will be shown.

Chapter 4

Results and Discussions

4.1 Introduction

In the previous chapter, the detailed explanation of our method has been provided. This chapter is aimed to show the comparative analysis of our proposed model. And other discussion will be indicated.

The model is implemented with python language of version 3.8. The program has been run on anaconda version 3 environment. The whole process has been conducted on Intel® Core^{TM} i7-8750H processor with 16 GB RAM. For the final implementation, healthcare heart study dataset is used [15]. It is a open source heart disease dataset.

4.2 Dataset Description

Healthcare heart study dataset is a open source dataset containing the risk factors of heart diseases [15]. The dataset is consisted of 5110 instances and 11 features. It contains categorical and numerical values. Some of the features are shown along with the feature types in table 4.1.

Some of the sample data are also shown in table 4.2.

Figure 4.1 shows the distribution of the data in the two dimensional space. It provides a meaningful insights into the data. There might be a linear distribution of the data. So, linear classifier may show a reasonable accuracy.

Table 4.1: Dataset sample variables and types

Features	Data Type
Sex	Categorical
Age	Integer
Marital Status	Categorical
Smoking Status	Categorical
Residence Status	Categorical
Prevalent Stroke	Categorical
Glucose Level	Categorical
Work Type	Categorical
Heart Disease	Categorical
BMI	Continuous numerical
Hypertension	Categorical

Table 4.2: Sample data of [15] heart study dataset

Sex	Age	Marital Status	Smoking Status	Residence Status	Heart Disease	BMI	Hypertension
1	39	Yes	Active	Urban	1	24	1
0	46	Yes	No	Rural	0	25	0
1	24	Yes	Active	Urban	0	25.5	0

4.3 Impact Analysis

None of us are risk free from hypertension. It may occur any stage of life. Most of the cases mass people don't pay heed to the hypertension. Because it doesn't show any severe problem at it's early stages. People take it as a normal health condition [2]. When they reach at the final stage of the hypertension, they could realize the fact. But it's too late. So, the proposed model would help the mass people to be conscious about hypertension. As there is not clinical test needed, people can check whenever they need.

4.4 Evaluation of the Proposed Method

In this section, the experimental results will be shown. It is a important section where validation result of the model is shown.

4.4.1 Feature Extraction with Correlation Analysis

For finding the important features for hypertension, a correlation is calculated with the hypertension stage feature.

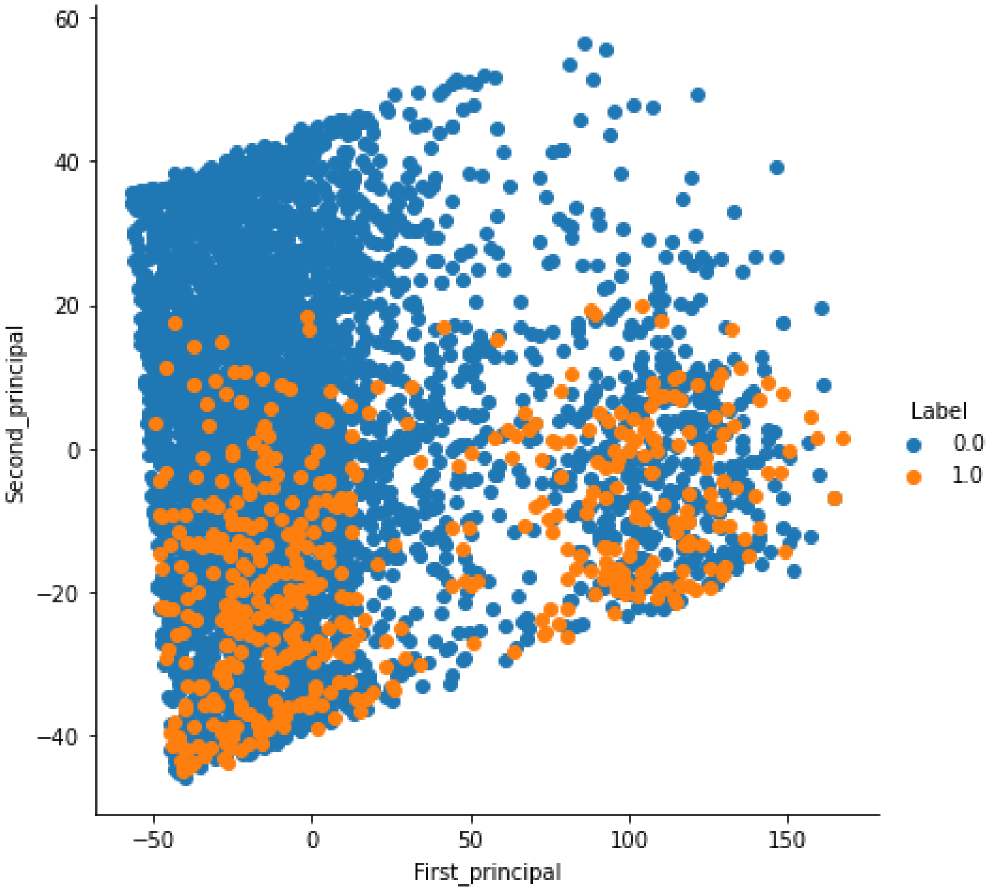

Figure 4.1: 2D distribution plot of the hypertension dataset.

Figure 4.2 indicates the heatmap of the features in terms of correlation. The following heatmap shows the correlation between the features with $(11X11)$ matrix. From the map we may find the correlation of hypertension stage with others features. All of the features are more or less correlated with hypertension stage. So, no column has been discarded.

4.5 Evaluation of Performance

Performance evaluation is one of the important task for a prediction model. It certifies a model whether it is functioning properly or not. For binary classification, it is easy to calculate the evaluation matrices. Our problem is a multi-class classification problem. Generally, precision, recall, f1-score and accuracy are some of the main evaluation matrices. It is worth mentioning that our data is imbalanced. In case of imbalanced data, weighted precision, weighted recall, weighted

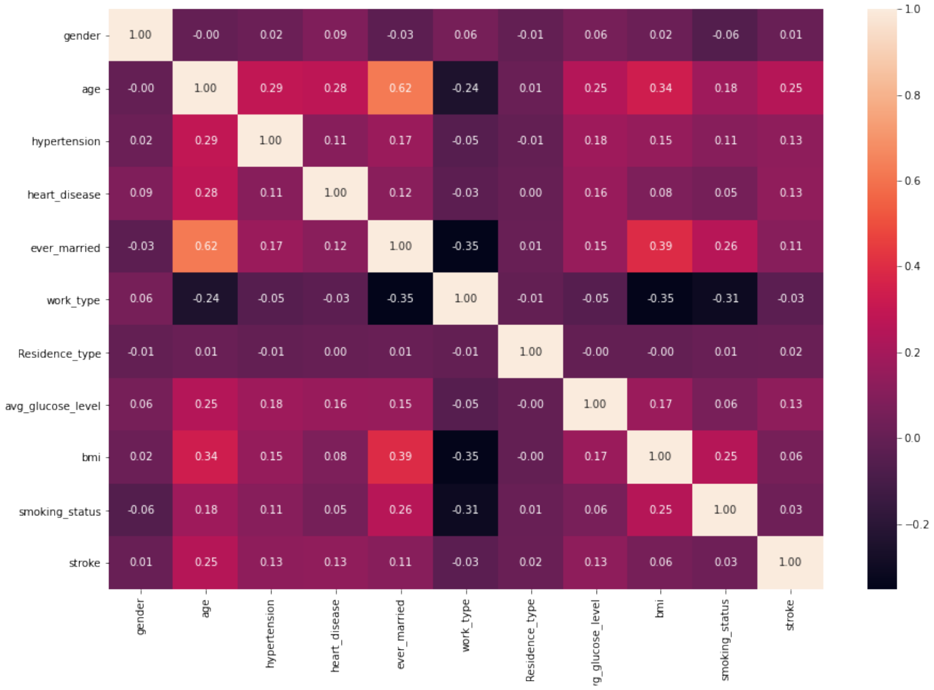

Figure 4.2: Correlation of the features with hypertension stage

f1-score and accuracy are calculated. The matrices are as follows-

- **Weighted Precision:** Precision indicates that how truly positive are the predicted positive.

$$Weighted - Precision = \frac{\sum_{i=1}^{n} S_i \frac{tp}{tp + fp}}{N} \tag{4.1}$$

  Where, n denotes the number of classes.

  N represents the total number of test instances.

  S_i represents the sample size of each class.

  tp stands for true positive.

  fp stands for false positive.

- **Weighted Recall:** Recall value indicates the proportion of actual positive class are correctly classified.

$$Weighted - Recall = \frac{\sum_{i=1}^{n} S_i \frac{tp}{tp + fn}}{N} \tag{4.2}$$

Where, n denotes the number of classes.

N represents the total number of test instances.

S_i represents the sample size of each class.

tp stands for true positive.

fn stands for false negative.

- **F1-score:** F1- score is the trade-off between precision and recall value.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.3}$$

- **Accuracy:** It indicates the big picture of evaluation matrices. Accuracy is the overall indication of overall performance.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{4.4}$$

Where, tp represents the value of true positive.

tn represents the value of true negative.

fp represents the value of false positive.

fn represents the value of false negative.

4.5.1 Evaluation of Performance of the Proposed Method

In this subsection, the weighted precision, weighted recall, f1-score and accuracy will be shown. The main purpose of using the concept of weighted precision and recall is evaluate the imbalance number of test sample of each classes.

Table 4.3: Evaluation matrices result for our proposed method.

Algorithm	Weighted Precision	Weighted Recall	F1-score	Accuracy
RF	0.82	0.91	0.86	0.91
DT	0.84	0.83	0.84	0.83
KNN	0.82	0.91	0.86	0.91
SVM	0.82	0.91	0.86	0.91
NB	0.88	0.85	0.86	0.85
LR	0.86	0.90	0.87	0.90

Table 4.3 shows the comparison of all the seven classifier algorithms. The algorithms are ANN, RF, DT, KNN, SVM, NB and LR. The comparative analysis clearly shows that random forest (RF), KNN and support vector machine (SVM) provides a good accuracy. Among the algorithms, RF is comparatively slow, KNN is also known as lazy algorithm because it has no training phase. So, finally the SVM classifier is selected as the predictive model of the proposed system. Figure 4.3 provides visual insights of the comparison of our proposed model.

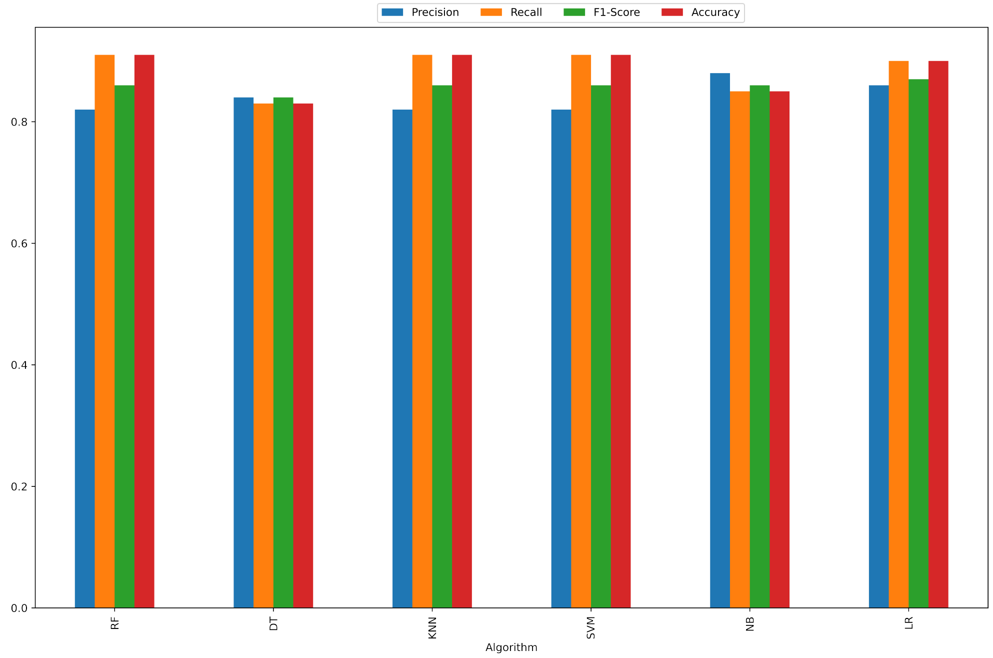

Figure 4.3: Evaluation matrices in terms of weighted precision, recall, f1-score and accuracy.

4.5.2 Sample Preliminary medical Suggestion for Hypertensive Patient

In this stage, the proposed model receives input data of the patients and firstly feed the data to the predictive machine learning model. The predictive model generates a output stage of hypertension of the user. A sample example of the suggestion is given below.

Input Data

1. Age - 60
2. BMI - 30
3. Sex - Male
4. Smoker - No
5. Heart Disease - Yes
6. Prevalent Stroke - Yes

The prediction model predicts the output hypertension for these inputs is high risk of hypertension. That means severe stage of hypertension. The suggestion for the data and predicted vale is given below.

Suggestion

1. Your target level Blood Pressure should be 147/90 mmHg. But it seems too high. As a aged person your cardiovascular risk is increased by >20 percent.
2. Reduce your BMI to normal range (18.5-24.9)kg/m^2. Balanced diet, increase your physical activity. Counselling, multi disciplinary program, medical and surgical therapy.Sudden cardiac death increases blood pressure.Lifestyle modification, physician's advice and supervised treatment is needed.Avoid sedentary lifestyle to get a healthy life.
3. Your hypertension condition is vulnerable. May be you are suffering with severe level of hypertension. Please, initiate therapy and go to hospital which facilitates proper treatment for hypertension.

4.6 Conclusion

In this chapter, the overall performance of the proposed method has discussed. The detail explanation of the dataset has been given. The process of feature extraction is also shown. Finally, the model has been evaluated with weighted precision , weighted recall, f1- score and accuracy. We have selected support vector machine (SVM) classifier as the best model. And then the sample suggestion has been shown.

Chapter 5

Conclusion

5.1 Conclusion

Millions of people around the world are suffering form cardiac disease. And hypertension is one of the most common cardiac diseases. Our proposed model would be a great assistant for mass people. With this system anyone can analyse his hypertension risk by simply inserting non clinical data. The model is able to provide an early indication of cardiac health risk. Additional feature of preliminary suggestion may someone helps to be cautious about the lifestyle so that the risk might be lessen. By using the system suggestion, anyone can take appointment for a doctor consistency. The model might be help to prevent hypertension risk to some extend.

Best model for the predictive system used is random forest. Because the model outperforms than any other model. Hope, the proposed system will add a new dimension to the medical sector.

5.2 Future Work

Medical science is a vast field where technology can be implemented for the betterment of mass people.

First of all, all the for pandemic situation, I have failed to collect enough data. Anyone can use the questionnaires and process to collect a bulk amount of data. So that the model can be trained with sufficient amounts of data. Obviously it will increase the accuracy and reliability of the model.

In case of suggestion, the proposed model has been used rule based structure. If a huge amount of text data can be collected, it is possible to generate suggestion dynamically with the natural language processing (NLP) techniques.

Medical image based hear disease analysis could be another addition to the work.

References

[1] *Hypertension*, `https://www.who.int/news-room/fact-sheets/detail/hypertension`, (Accessed on 04/14/2021) (cit. on p. 1).

[2] *Who | a global brief on hypertension*, `https://www.who.int/cardiovascular_diseases/publications/global_brief_hypertension/en/`, (Accessed on 04/14/2021) (cit. on pp. 1, 8, 20).

[3] D. M. Reboussin, N. B. Allen, M. E. Griswold, E. Guallar, Y. Hong, D. T. Lackland, E. (R. Miller III, T. Polonsky, A. M. Thompson-Paul and S. Vupputuri, 'Systematic review for the 2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A report of the american college of cardiology/american heart association task force on clinical practice guidelines,' *Hypertension*, vol. 71, no. 6, e116–e135, 2018 (cit. on pp. 1, 8).

[4] J. Y. Islam, M. M. Zaman, S. A. Haq, S. Ahmed and Z. Al-Quadir, 'Epidemiology of hypertension among bangladeshi adults using the 2017 acc/aha hypertension clinical practice guidelines and joint national committee 7 guidelines,' *Journal of human hypertension*, vol. 32, no. 10, pp. 668–680, 2018 (cit. on pp. 1, 6).

[5] A. M. Islam and A. A. Majumder, 'Hypertension in bangladesh: A review,' *Indian heart journal*, vol. 64, no. 3, pp. 319–323, (cit. on pp. 1, 8).

[6] *High blood pressure (hypertension) - symptoms and causes - mayo clinic*, `https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410`, (Accessed on 04/14/2021) (cit. on pp. 4, 8).

[7] E. Ritz, C. Strumpf, F. Katz, A. Wing and E. Quellhorst, 'Hypertension and cardiovascular risk factors in hemodialyzed diabetic patients.,' *Hypertension*, vol. 7, no. 6_pt_2, p. II118, 1985 (cit. on pp. 6, 8).

[8] P. M. Kearney, M. Whelton, K. Reynolds, P. Muntner, P. K. Whelton and J. He, 'Global burden of hypertension: Analysis of worldwide data,' *The lancet*, vol. 365, no. 9455, pp. 217–223, 2005 (cit. on p. 6).

[9] S. Mohan, C. Thirumalai and G. Srivastava, 'Effective heart disease prediction using hybrid machine learning techniques,' *IEEE Access*, vol. 7, pp. 81 542–81 554, 2019. DOI: `10.1109/ACCESS.2019.2923707` (cit. on p. 6).

[10] A. Dinh, S. Miertschin, A. Young and S. D. Mohanty, 'A data-driven approach to predicting diabetes and cardiovascular disease with machine learning,' *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–15, 2019 (cit. on p. 7).

[11] S. Sakr, R. Elshawi, A. Ahmed, W. T. Qureshi, C. Brawner, S. Keteyian, M. J. Blaha and M. H. Al-Mallah, 'Using machine learning on cardiorespiratory fitness data for predicting hypertension: The henry ford exercise testing (fit) project,' *PLoS One*, vol. 13, no. 4, e0195344, 2018 (cit. on p. 7).

[12] H. F. Golino, L. S. d. B. Amaral, S. F. P. Duarte, C. M. A. Gomes, T. d. J. Soares, L. A. d. Reis and J. Santos, 'Predicting increased blood pressure using machine learning,' *Journal of obesity*, vol. 2014, 2014 (cit. on pp. 7, 8).

[13] D. LaFreniere, F. Zulkernine, D. Barber and K. Martin, 'Using machine learning to predict hypertension from a clinical dataset,' in *2016 IEEE symposium series on computational intelligence (SSCI)*, IEEE, 2016, pp. 1–7 (cit. on p. 7).

[14] T. J. Dawes, A. de Marvao, W. Shi, T. Fletcher, G. M. Watson, J. Wharton, C. J. Rhodes, L. S. Howard, J. S. R. Gibbs, D. Rueckert *et al.*, 'Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: A cardiac mr imaging study,' *Radiology*, vol. 283, no. 2, pp. 381–390, 2017 (cit. on p. 7).

[15] *Stroke prediction dataset | kaggle*, `https://www.kaggle.com/fedesoriano/stroke-prediction-dataset`, (Accessed on 04/29/2021) (cit. on pp. 8, 16, 19, 20).

[16] O. A. Carretero and S. Oparil, 'Essential hypertension: Part i: Definition and etiology,' *Circulation*, vol. 101, no. 3, pp. 329–335, 2000 (cit. on pp. 8, 12).

[17] M. A. Hall, 'Correlation-based feature selection for machine learning,' 1999 (cit. on p. 12).

[18] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein and M. Klein, *Logistic regression.* Springer, 2002 (cit. on p. 14).

[19] L. Kozma, 'K nearest neighbors algorithm (knn),' *Helsinki University of Technology*, 2008 (cit. on p. 14).

[20] I. Rish *et al.*, 'An empirical study of the naive bayes classifier,' in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, 2001, pp. 41–46 (cit. on p. 14).

[21] W. S. Noble, 'What is a support vector machine?' *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006 (cit. on p. 15).

[22] Y.-Y. Song and L. Ying, 'Decision tree methods: Applications for classification and prediction,' *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015 (cit. on p. 15).

[23] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, 'Random forest: A classification and regression tool for compound

classification and qsar modeling,’ *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003 (cit. on p. 15).

[24] B. Yegnanarayana, *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009 (cit. on p. 15).

[25] Y. Bao and Z. Liu, ‘A fast grid search method in support vector regression forecasting time series,’ in *International Conference on Intelligent Data Engineering and Automated Learning*, Springer, 2006, pp. 504–511 (cit. on p. 17).

[26] J. E. Hall and M. E. Hall, *Guyton and Hall textbook of medical physiology e-Book*. Elsevier Health Sciences, 2020 (cit. on p. 17).