

THALASSEMIA PREDICTION USING A MACHINE LEARNING APPROACH

BY
PUSHPITA KARMAKER
ID: 181-15-10908

ANANYNA DEVANATH
ID: 181-15-11327

SHAHNAZ AKTER
ID: 181-15-11032

This Report Presented in Partial Fulfillment of the Requirements for The Degree of
Bachelor of Science in Computer Science and Engineering

Supervised By

ABDUS SATTAR
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

MD. RIAZUR RAHMAN
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2022

APPROVAL

This Project titled “**Thalassemia Prediction Using A Machine Learning Approach**”, submitted by **Pushpita Karmaker, Ananyana Devanath, Shahnaz Akter** ID No: **181-15-10908, 181-15-11327, 181-15-11032** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 3 January 2022.

BOARD OF EXAMINERS



Dr. S.M Aminul Haque

Associate Professor and Associate Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Chairman



Naznin Sultana

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Raja Tariqul Hasan Tusher

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Dr. Dewan Md. Farid

Professor

Department of Computer Science and Engineering

United International University

External Examiner

DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Abdus Sattar**, Assistant Professor, and Department of CSE Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

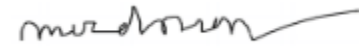
Supervised by:



Abdus Sattar

Assistant Professor
Department of CSE
Daffodil International University

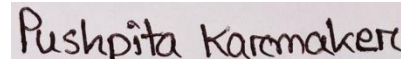
CO-SUPERVISED BY:



Mr. Md. Riazur Rahman

Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:



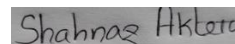
Pushpita Karmakar

ID: 181-15-10908
Department of CSE
Daffodil International University



Ananyana Devanath

ID: 181-15-11327
Department of CSE
Daffodil International University



Shahnaz Akter

ID: 181-15-11032
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First and foremost, we express our profound thankfulness to almighty God for his divine mercy, which has enabled us to complete the final thesis successfully.

Abdus Sattar, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University, Dhaka, has been extremely helpful. Our supervisor was able to finish this thesis because of his extensive knowledge and intense interest in the subject of "Machine Learning." His unwavering patience, intellectual direction, constant encouragement, frequent and energetic supervision, constructive criticism, helpful recommendations, and reading and editing several inferior versions at all levels allowed me to complete this thesis.

Professor Dr. Touhid Bhuiyan, Professor and Head, Department of CSE, as well as the other academic members and personnel of Daffodil International University's CSE department, deserve our heartfelt gratitude for their contributions to the completion of our thesis.

We would like to thank everyone of our Daffodil International University classmates who took part in this conversation as part of their course work.

Finally, we must recognize and thank our parents for their constant support and dedication.

ABSTRACT

Thalassemia is one kind of genetic blood disease/disorder which is caused if human body can't produce sufficient hemoglobin. It is known that hemoglobin is a very common essential part of anyone's body. RBC's of human body don't work efficiently if there is any lacking of hemoglobin. Then a little amount of healthy RBC's travel in one's bloodstream. Oxygen which is carried by red blood cell is kind of food, that food cells can utilize to work. Due to lacking of sufficient healthy RBC's, sufficient oxygen can't be delivered to every cells of the body, which can be a reason to cause a person to anemia, that is responsible to damage organs and lead one to death. In this research we are working about predicting the existence of Thalassemia with ML, an important part of AI. We implemented very popular ML algorithms on our processed dataset. We used k-nearest neighbor (*k*NN), logistic regression, support vector machine (SVM), naïve Bayes, random forest, adaptive boosting (ADA boosting), XGBoost, decision tree, multilayer perception (MLP) and gradient boosting classifier. In our work, out of ten algorithms, ADA BOOST algorithm gave the greatest output which is related on accuracy and it was 100%.

TABLE OF CONTENTS

CONTENTS

	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv

CHAPTER

CHAPTER 1: INTRODUCTION 1-4

1.1 Introduction	01
1.2 Motivation	02
1.3 Problem Definition	02
1.4 Research Question	02
1.5 Research Methodology	03
1.6 Research Objective	03
1.7 Article Layout	3-4

CHAPTER 2: BACKGROUND STUDY 5-18

2.1 Introduction	5
2.2 Related work	5-12
2.3 Literature review summary	13-17
2.4 Bangladesh Perspective	17
2.5 Challenges	18

CHAPTER 3: RESEARCH METHODOLOGY 19-30

3.1 Introduction	19
3.2 Procedures of data collection	19-20
3.3 Application of ‘SMOTE’ technique	21
3.4 Research subject and instrumentation	21
3.4.1 Proposed Methodology	22
3.4.2 Data Processing	22-24
3.5 Statistical Analysis	25-30

CHAPTER 4: PERFORMANCE OF THE PROPOSED MODEL	31-46
4.1 Introduction	31
4.2 Results and Analysis of Experiments	31
4.2.1 Experimentation	31
4.2.2 Analytical Methodology	32 -46
CHAPTER 5: RESULT COMPARISON AND Analysis	47-57
5.1 Project interface	47-55
5.2 Project Architecture	55
5.3 Impact on society	56
5.4 Impact on Environment	56
5.5 Ethical Aspects	56
5.6 Sustain ability plan	57
CHAPTER 6: CONCLUSION AND FUTURE WORK	58-59
6.1 Summery of the research	58
6.2 Limitations and conclusions	58
6.3 Implication for further study	59
REFERENCES	59-61
APPENDIX	62

LIST OF FIGURES

FIGURES	PAGE NO.
Figure 3.2.1 Collected Dataset demo	20
Figure 3.4.1 Proposed Methodology	22
Figure 3.4.2 Data preprocessing	23
Figure 3.5.1 Pie chart before using 'SMOTE' technique	25
Figure 3.5.2 Pie chart after using 'SMOTE' technique	25
Figure 3.5.3 Scatter plot (PLT VS HGB)	26
Figure 3.5.4 Scatter plot (TLL,MCHC,RDW)	
Figure 3.5.5 Scatter plot (RBC vs AGE)	27
Figure 3.5.6 Histogram (HGB vs THALASSEMIA)	27
Figure 3.5.7 Histogram (PCV vs THALASSEMIA)	28
Figure 3.5.8 Histogram (SEX vs THALASSEMIA)	28
Figure 3.5.9 Histogram (AGE vs THALASSEMIA)	29
Figure 3.5.10 Correlation Matrix	29
Figure 3.5.11 Parallel Co-ordinates	30

FIGURES	PAGE NO.
Figure 4.1 ROC curve of LR algorithm	33
Figure 4.2 ROC curve of KNN algorithm	33
Figure 4.3 ROC curve of SVM algorithm	34
Figure 4.4 ROC curve of DT algorithm	34
Figure 4.5 ROC curve of RF algorithm	35
Figure 4.6 ROC curve of NB algorithm	36
Figure 4.7 ROC curve of XG-BOOST algorithm	36
Figure 4.8 ROC curve of ADA-BOOST algorithm	36
Figure 4.9 ROC curve of MLP algorithm	37
Figure 4.10 ROC curve of STOCHASTIC	38
Figure 4.11 Confusion matrix LR algorithm	39
Figure 4.12 Confusion matrix KNN algorithm	40
Figure 4.13 Confusion matrix SVM algorithm	40
Figure 4.14 Confusion matrix DT algorithm	41
Figure 4.15 Confusion matrix RF algorithm	42
Figure 4.16 Confusion matrix NB algorithm	42
Figure 4.17 Confusion matrix XG-BOOST algorithm	43
Figure 4.18 Confusion matrix ADA-BOOST algorithm	44
Figure 4.19 Confusion matrix MLP algorithm	45
Figure 4.20 Confusion matrix GRADIENT-BOOST algorithm	46

LIST OF TABLES

TABLE NO.	PAGE NO.
Table 2.1 Literature review summary	13
Table 5.1 Classifier performance evaluation	47
Table 5.2 Comparative analysis	48

CHAPTER 1

INTRODUCTION

1.1 Introduction

Thalassemia is the name of inborn disorder that affect Hemoglobin's, substances in the blood. Thalassemia patients can produce too little Hb (Hemoglobin), which is utilized by RBC's to provide oxygen around the whole body. This disorder often can cause anemia. Mainly People who are from Mediteranean, Asian (South and Southeast) and Middle Eastern are seriously influenced by it. Alpha and Beta are most important types of Thalassemia, here Beta Thalassemia is more dangerous than alpha thalassemia. If anyone is beta Thalassemia carrier he/she may not have any health issues, but he/she is at risk of having baby with Thalassemia. At present this disease has become a global public health issue. Asia is the hotspot of Hemoglobinopathis and it's near about twenty three percentage of the world. Thalassemia related informations significantly in Asia (south) comes from research in India. An imbalanced frequency of beta Thalassemia heterozygote, carrier in range of 1 -10% has reported in various parts of India. Finally prevalence of beta Thalassemia was between 2.78 and 4% in India. Bangladesh is one of the most overpopulated countries in this world. Beta Thalassemia and HBE variant are also found in Bangladesh increasingly. Per year around 1 lakh children are born with this dangerous disease in the world. This is not a contagious disease, it is inherited from parents. Even a study shows that, around 25% of babies born with thalassemia are affected by the thalassemia gene in both parents. Minor thalassemia may not effect dangerously but thalassemia major needs effective treatment. Blood transfusion in regular basis is a common treatment for thalassemia major. Again bone marrow transplantation is also an important treatment for thalassemia. Prediction is required for one's upcoming treatment.

We all know that ML is outcome of AI, permits software apps to be more faultless at determination results except explicitly programmed to do so. Algorithms of ML use ancient data as input to analyze new output values. SO, we have used machine learning for this Thalassemia prediction.

1.2 Motivation

Severe thalassemia causes early death to patients. Even many types of complications are occurred due to major thalassemia. A research shows that most of the thalassemia patients may live up to 25 to 30 years of their lifetime, if we can provide them sufficient facilities that will help the patients live up to the age 60. So proper medication can increase life expectancy of a thalassemia patient. If a thalassemia patient left untreated, that will cause issues in the liver, heart, and spleen. Life-threatening issues of Thalassemia in children are infections and heart failure. Frequent blood transfusion is needed for both adult and children if they are suffering from major thalassemia. Otherwise, for major thalassemia, the following complications can occur like bone deformities, growth dilation, facial changes. It makes one's bone marrow expand, at that time one's bones become widen. So that can be an abnormal bone structure. Thalassemia patient doesn't have a simple expectancy of life. heart complications coming from beta thalassemia major can build this situation fatal before 30 years old. So prediction will be helpful for their early treatment. We need to keep a special focus so that no patient should die untreated. We haven't noticed much research in this area. We are going to do this analysis implementing ML approaches.

1.3 Problem Definition

We have applied multiple types of machine learning approaches to predict Thalassemia with the expectation of higher accuracy. ML's main concept is to study of some implementing algorithms and data permits one's computers to do tasks except instructions and human user's input. ML is used by computers performing auto improvement using actual-world examples and data to do so, rather than human input. Image or Speech recognition, Medical diagnosis, Statistical issues, Predictive analytics, Extraction are the main categories of ML applications. ML is interesting, amusing as programs learn from examples. From the data that we have gathered, a ML technique can automatically analyze and learn the format resident in that data for providing a solution to the problem we are trying to solve. After thinking about all we have used that techniques.

1.4 Research Questions

- What will be our proper executable dataset?
- How do we identify expected patients?
- How machine learning is applied to get expected outcomes?
- Is there any other way to identify Thalassemia patients?
- How we can make trained out our main data to the ML model?
- How much amount of data do we have in our collection?
- Does our gathered data and ML will be compatible? Which new techniques of machine learning should be applicable?

1.5 Research Methodology

In this part of our research paper, we used an Experiment Data Set, Pre-processing of data, Model Development several types of graphs analysis, testing dataset. Finally, the performance of the proposed model will be narrated at the end of our chapter.

1.6 Research Objectives

We expect from our research that will help people to predict Thalassemia with more accuracy. Applying this approach, people may know about health condition more quickly and easily. People will also learn more about analysis with ML. Successful implementation of existing or new ML algorithms for prediction Thalassemia. Predictions can make aware about Thalassemia to the patients family to be more careful about it. Severe patients requires very extra care and medication otherwise it will cause a significant harm. It will protect us and our society from the adverse effects of untreated Thalassemia. Again doctors can make a collaboration with our model which will assist them to identify Thalassemia positive or negative people. On other hand, the development of a data set for Thalassemia in the context of Bangladesh. Publication of one or more articles in conference internationally proceedings or journals.

1.7 Article Design

The schedule form of the contents of our dissertation:

- Canton 1 narrates the survey's initiation, motivation of research, rationale of the study, questions about research, and inevitable motive.
- Part 2 contains works that are linked, research summary, the situation's magnitude, and ultimatum.
- Item 3 gives brief overview of the workflow of this research, procedures of specifics bevy, and a quantitative investigation and feature enactment.
- Volume 4 accords a snapshot of avant-garde perusal and appraisal, but also some pertinent debates, along with the cramming unearthing in numerical and graphic form..
- Portion 5 addresses the repercussions of this scrutiny..
- Paragraph 6 purvey an encapsulation of this study's uncovering, as well as curb and posterior research.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

Here we are talking about few relevant topics, research summary, the fact of the problem, and challenges in this section. We have summarized some research papers, related works which are relevant for our work in the segment of related work.

We have prepared a outline of some related works and dispose them in a table for easy understanding for us. The scope of the problem part demonstrate how we can give our contribution to the issue with our work model. At the end, the Challenges part accommodate some words about the obstacle and dangers we experience during the this journey.

2.2 Related Works

In this part of this paper going to show the related works which were done in past by other researchers on Thalassemia forecasting. We have analyzed and specifically studied the efforts of the researchers to acknowledge the processes and methods which were shown by them.

Eyad H. Elshamiet al.[1] has given a proposal of a system for investigation for thalassemia forecasting ,that will using data mining classifiers which depends on CBC. His paper used NB, DT and Neural Network as its algorithm & got accuracy in naive Bayes (93.7%), decision tree (93.64%) and Neural Network (95.71%). Three data mining classifiers were used in this investigation. Each of the classifiers are being used to determine differences between thalassemia patients- with its other levels-: normal persons, iron deficiency patients and the patient who belongs to other blood diseases. The results suggest that more investigation is required to determine that if the classification classes were reduced into Thalassemia trait, iron deficiency, and normal, then examine if the results can provide us with more accurate results.

NgoziChidozieEgejuru et al.[2] has proposed a study titled which can tell the danger of thalassemia in all age peoples based on identified risk of thalassemia. Thalassemia risk factor's knowledge was for identifying st interview with medical personel and questionnaire .These were used for collection empirical database of medical on the parameters. Supervised ML algorithms

was used for formulating the predictive model for danger of thalassemia using the identified, collected parameters and data. The model of prediction for the risk of thalassemia was calculated using the Waikato Environment for Knowledge Analysis (WEKA). The model which can proceed simulation was validated by implementing the historical data which are collected from the hospitals discussing the parameters and the danger of Thalassemia. This model used WEKA (software decision tree based model), Naïve Bayes, multilayer perceptron as algorithm. The accuracy is multilayer perceptron (100%), Naïve Bayes (94.12%).

MunaQais Mohammed et al.[3] has proposed a study, In his research we get a survey of various techniques based on AI for the classification and detection of thalassemia using the (parameter) of the CBC test which include RBC, MCV, HGB, HTC, HB to distinguish thalassemia minor alpha and thalassemia major beta patients. NB, DT, SVM, and neural network classification approaches are used. Observation of this survey that certain ML algorithms, such as DT, SVM, ANN, and KNN, gives us better accuracy. Moreover, these variety of algorithms behave in different way related on different different factors. Finding better prediction results, situation consideration yet most significantly the dataset and feature selection. Additionally, this paper includes a survey of various types of ML methods using by various researchers, with each ML technique relying on the datasets and characteristics.

Yi-Kai Fu et al. [4] published a paper, At this a classifier for determining the thalassemia positive or negative microcytic anemia was created through combination of exciting indices with ML approach. 350 patients, age over 40 whose anemia diagnosis, hemoglobin gene profiles and cbc and were significantly reviewed. 13 prior established indices were used to current cohort and the specificity, sensitivity, negative and positive predictive values were for calculation. A SVM curve for the sampled datasets used Monti-Karlo-cross validation. The model showed performance average AUC of 0.76 and avg error of 0.26.

A paper had already been submitted by Fatemeh Yousefian et al. [5]. The used tools for screening thalassemia are researched and appraised in this paper. The accessible factors in the CBC test in patients with thalassemia diagnosis include HGB, RBC, MCV, and HTC, among which HGB, RBC, MCV, and HTC have a substantial effect on the disease diagnosis. The

patient with thalassemia is identified based on available values from automation and CBC results. Artificial intelligence algorithms are utilized to appropriately interpret laboratory data, resulting in increased accuracy in disease diagnosis, which has a major effect on the treatment process and patient health improvement. Various publications applied the KNN, NN, MLP, SVM, and DT procedures, with accuracy varying from 93.2 % to 94.35 percent, MLP to 99.12 %, NN to 95.71 %, and DT to 95.71 percent of the total (93.64 percent).

By integrating clinical information of Hypochromic microcytic anemia discovered in grownups, V. Laengsri et al. [6] proposed a system to determine identifying thalassemia. A discriminant model was built using five ml algorithms: decision tree, kNN, genetic approaches (ANN), random forest (RF), and SVM classifiers (Support vector). In the work separating genetic component and anemia," performance was evaluated concerning respect to 13 discriminant formulas and indices. A total of 186 individuals' data were enrolled. The RF model's interpretable rules were proposed to demonstrate the use of a mixture of RBC indices to discern Disease from Target text. An Proposed technique based on 7 RBC boundaries was used to create the web-based tool 'ThalPred.' ThalPred attained external accuracy, MCC, and AUC of 95.59, 0.87, and 0.98, correspondingly, in its prediction findings.

Roberta Risoluti et al. [7], has been published. 16 thalassemia transnational ventilator patients, 18 thalassemia mutualists pre sick people, and 14 thalassemia major patients were among the 128 participants in this study. Positive patients were seen to be clearly distinguishable from tissue samples due to a difference in thermal behavior. Chemometric analysis identifies changes in blood composition, and a model for -thalassemia forecasting was established and validated to discriminate all individuals. The TGA/Chemometric system also permitted for the distinction of thalassemia patients due to the seriousness of anemia, whereas indices and CBC were unable to identify TI-NTD, TI-TD, and TM-TD patients at the first level test. The Thermogravimetric metric was reliably used to treat thalassemia, with a 100percent of overall validation set frequency. When compared to mean corpuscular volume and mean corpuscular hemoglobin, chemometric research suggests indicate hemoglobin, red blood cell count width, and Cbc are the clinical markers in thalassemia (HB). For a modification of thalassemia categorization, new insights into the significance of hematological characteristics were offered.

The RF algorithm Technique was used in a work by F R Aszhari et al. The results show that using multiple five in the range of 70percentage to 85percent in terms of sequence data, the random forest algorithm can provide the best precision, accuracy, and recall of 100 percent. They believe that further research will be able to develop a new method for modifying this method to predict or classify other diseases. They hope that by participating in this study, the results will aid the medical community in grouping and order to detect whether a patient has thalassemia whether if, so that the patient can receive the appropriate treatment to extend their life expectancy and reduce the likelihood of Thalassemia in future generations.

Alaa S. AlAgha et al. [9] suggested a framework. There are two basic stages in this proposed paradigm. To begin, a balancing technique known as SMOTE is postulated also used to address the problem of a highly imbalanced class distribution in a dataset. In the second step, four classification models are utilized to detect a variety people and those who have -thalassemia: k-nearest neighbors, naive Bayesian, decision tree, and deep convolutional neural channel. The findings show that using the SMOTE to spread may be achieved. The observation revealed that at a 400 percent frame interpolation SMOTE proportion, the NB classifier had the best performance in characterizing across ordinary and carriers. This combo has a 99.47% clarity and a 98.81 % acuity.

Monalisha Saikia Borah et al. [10] developed a scheme that can predict hemoglobin variants. Algorithms involve logistic regression, support vector classifier, perceptron classifier, Normal distribution Nave Bayes, linear SVC, multi-layer feedforward neural gradient descent, random forest, and decision tree. Expertise was 93.89 percentage, recalling was 92.78 %, and the f1-score was 93.33 percent. Because it aids in the prediction and diagnosis of Hb variations, this work demonstrates that machine learning is useful in healthcare. Machine learning, which employs a variety of classifiers, can be used to predict certain types of Hb polymorphisms. Random forest and decision tree classifiers were found to be the front-runners in terms of classifying Hb variations in assays. In brief inquiry, as well as a variety of data, may result in higher "quality," "vaguely remember," and "f1-score" scores, all of which signify improved consistency. In the future, better data mining methods may be created, allowing for a more real study.

Dr. C.A.D.M.N.C. Kolambage et al. [11] has proposed a model in the paper named "Design, Development and Implementation of the Quantitative Simulation Toolkit ml toolkit Predicted using Full Blood Count Indices and Hb Derivatives, Thalassemia Delivery status can be resolved.. "In this paper which is used algorithms are an ANN, Random Forest Algorithm with four hidden layers. Hypothesis 1 got taught how to the distinguish in both normal persons and sufferers of Hemoglobin synthesis, as well as the Random Forest outputer generated the ANN Effectiveness 93.5 and the F1 score of 87.5 percent. Model 2 differentiate between Alpha thalassemia silent carrier and Alpha thalassemia traits, and the RF again showed performance with accuracy 88.8 and an F1 score of 86.3 percent. Model 3, which will be developed by the The gap is thalassemia deficiency, breeders, and physically active. On model 3, ANN surpassed the competition with an exactness and True positive rate of 83.3 percentile. To ensure that results obtained in this case are reproduced in the real world, the author suggests the following process to be followed in the model deployment. The built model can be combined with the successful models described above to develop a machine learning based investigation algorithm that saves money and time. Applying the approach in this findings should be used to build a picture. identify beta thalassemia carrier state which is based on blood counted alone with an important future worked that can convey. The considerable with burden that country is facing in non-communicable diseases such as mental health, diabetes heart disease, machine learning tools are developed for prognostic, diagnostic tasks have the potential of providing support in this regarded.

WaranyuWongseree et al. [12] offered a system .In this paper used algorithm is GP-based decision tree: STROGANOFF.Firstly case study with 10 classes and 13 input features, the avg classification accuracy when testing is 90%.Secondly case study with 15 classes and 15 input features, the avg classification accuracy during testing is 82%. The performance of the GP-based DT is comparable to that of the multilayer perceptron with one hidden layer,An elimination of redundant input features may not affect the classification result of the multilayer perceptron which can be tested and improve further.

BetülÇil, HakanAyyıldız et al. [13] offered a model in the paper And the used algorithms are SVM, KNN, Logistic Regression, Extreme Learning Machine ,Extrem Learning Machine classification algorithms. Percentages of accuracy are 96.30% for female, 94.37% for male, and

95.59% in co-evaluation of male and female patients also which were obtained. The offered system could be used to differentiate IDA and β -thalassemia. Again this system could be easily applied to deal with the biochemistry issues. It may be concluded that the present case provides a faster and less costly solution for differentiation of anemia and β -thalassemia.

Paokanta, Patcharaporn, et al. [14] has proposed a system in the paper "The study provided a comparison of the classification performance of ML techniques that use PCA to screen genotypes in β -thalassemia patients. The future of this study is to reduce the dimension of data before classification. According to the PCA-the result of the method and classification method is that if the percentage of accuracy per K reaches 86.61, the percentage of nearest neighbor (KNN), naive bays, basic network (BN), and polynomial logistic regression accuracy is 85.83. Indicates that the MLP is the best algorithm when required. 85.04, 85.04, 82.68% were respectively.

E.R. Susanto et al. [15] proposes a paper entitled "Implementing a Fuzzy-Based Model for Prediction of Thalassemia Diseases". In this paper, they hope to develop a new model for predicting thalassemia in children. This model uses fuzzy-based rules. The novelty of this article has shown that our model has four outputs: thalassemia major, intermediate, minor, and non-thalassemia. In the previous article, there are only three outputs. In this case, they plan to implement architecture to use a fuzzy-based approach disease based on CBC data.

Farhadi et al. [16] The publication, "Predicting Transfusion Complications in Thalassemia Patients Using Deep Learning Methods", proposed a system that uses machine learning algorithms to predict the risk of post-transfusion complications in positive patients. The study used machine learning to predict 4,444 post-transfusion complications patients. After all, the deep learning approach provided the most accurate predictions. This method is used to identify patients who have experienced complications prior to transfusion. Appropriate alternatives can be used to treat these patients and prevent or reduce transfusion complications. In cross-section, data on 3489 cases from 12 Thalassemia Centers in Tehran Province and 14 Thalassemia Centers in Mazandaran were collected in 2018. Various classification models were trained and investigated in this dataset, including classical and deep learning techniques. The results show that machine learning methods show excellent accuracy in predicting the risk of post-transfusion complications. Therefore, the deep learning method significantly compared to others (accuracy = 0.21, sensitivity = 0.77, f1score = 0.33).

P. Paokanta et al. [17] published a paper entitled "Data Type Efficiency for Identifying the Performance of Machine Learning Methods for Screening β -Thalassemias" and determined the optimal data type for each method. Genotypes of β -thalassemia patients are classified using β -thalassemia data. The results of this suggest that the data type is a nominal scale that can be used for Bayesian networks (BN) and multinomial logistic regression with accuracy of 85.83% and 84.25%, respectively. In addition, data types such as interval scales can be effectively used for KNearest Neighbors (KNN), MultiLayer perceptron, and Nave Bayes with 88.98, 87.40%, and 84.25% accuracy, respectively.

Sacco et al. [18] proposed an analysis in a paper titled "Random forest analysis: a new approach to beta-thalassemic classification". presence of two or more clusters (TDT vs. NTDT), to propose a new classification system for beta-thalassemia in thalassemia syndromes.

Jahangiri, Mina et al. [19] proposed a model in a study, first time a tree-based algorithm has been used to distinguish between TT and IDA. A total of 144 hypopigmental microcytic anemias between the ages of 18 and 40 were recruited from Ayat Hospital in Tehran. To distinguish between diagnoses, researchers have classified and regression trees, chi-square automatic interaction detectors (CHAID), exhaustive chi-square automatic interaction detectors (ECHAID), unbiased, rapid, QUEST, biased. Classification rules (CRUISE) is detection and estimation. Mean corpuscular volume (MCV) was determined as the primary predictor of discrimination. Differential diagnosis of TT by IDA and all the tree-based approaches discussed found appropriate specificity, sensitivity, Youden index, accuracy, false positive and negative rates, positive and negative predictors, and AUC. rice field. In the area below the 0.99 curve, classification rules using unbiased interaction selection and estimation showed more accurate classification.

2.3 Literature Review Summery

Many efforts has already been taking place in the field of prediction and for detection with ML algorithm and approaches of data mining. These days the application of ML technology has rapidly increased for these various disease detection. In this section, the comparisons between these related works have shown.

Here are the comparison of different thesis works with their subject, methodology, and the outcome are given below in Table 2.1.

Article	Author name	Used Algorithm	Prediction	Accuracy
[1]	Eyad H. Elshami	Decision Tree, Naïve Bayes, Neural Network	Diagnosis of Thalassemia	Decisiontree (93.64%), naive Bayes (93.7%), and Neural Network (95.71%)
[2]	NgoziChidozieEgejuru	WEKA (software decision tree based model), Naïve Bayes, multilayer perceptron	Thalassemia risk prediction	Naïve Bayes (94.12%), multilayer perceptron (100%).
[3]	MunaQais Mohammed	Decision tree, support vector machine (SVM), Naive Bayes, and ANN, neural network KNN, ANN.	thalassemia detection and classification	Not mentioned
[4]	Yi-Kai Fu	vector machine learning (SVM)	Differentiating Thalassemia and Non-	SVM 95%.

			Thalassemia Patients	
[5]	FatemehYousefian	KNN, MLP, NN, DT and SVM	Prediction Thalassemia	KNN (93.2%), MLP (99.12%), DT (93.64%), NB (94.35%), NN (95.71%).
[6]	V. Laengsri	random forest (RF), artificial neural network (ANN), k-NN, decision tree (DT), and support vector machine learning (SVM)	discriminating thalassemia trait and iron deficiency anemia	Got High accuracy in (RF, SVM), Poor accuracy in (KNN, DT).
[7]	Roberta Risoluti	TGA/chemo metrics method	Update onthalassemiadiagnosis	got full accuracy.
[8]	F R Aszhari	Random forest(RF)	Classification of thalassemia	70-85%
[9]	Alaa S. AlAgha		Identifying β -thalassemia	
[10]	MonalishaSaikia Borah	Logistic regression, k-nearest neighbor (KNN), support vector classifier (SVC), Gaussian Naïve Bayes,	Predicting Hemoglobin Variants	Not mention

		perceptron classifier, linear SVC, decision tree, stochastic gradient descent, random forest, and multi-layer perceptron		
[11]	Dr. C.A.D.M.N.C. Kolambage	Random Forest, Artificial Neural Network (ANN)	Predictive Modelling Tool to Accurately Predict Thalassemia Carrier state using Full Blood Count Indices and Hemoglobin Variant.	
[12]	WaranyuWongseree	GP-based decision tree: STROGANOFF	Thalassemia classification	
[13]	BetülÇil, HakanAyyıldız	Logistic Regression, KNearest Neighbors, Support Vector Machine, Extreme Learning Machine and Regularized Extreme	Discrimination of β -Thalassemia and Iron Deficiency of Anemia	94.37% for male, 96.30% for female, and 95.59% in co-evaluation of male and female

		Learning Machine classifications		
[14]	Paokanta, Patcharaporn,	The Multi-Layer Perceptron (MLP). K-Nearest Neighbors (KNN), Naive Bayes, Bayesian Networks (BNs) and Multinomial Logistic Regression	The Knowledge Discovery of β -Thalassemia	MLP 86.61%, KNN 85.83%, Naive Bayes 85.04%, Bayesian Networks 85.04% and Multinomial Logistic Regression 82.68%
[15]	E. R. Susanto	Fuzzy based model	Prediction of Thalassemia Diseases	
[16]	Farhadi		The Prediction of Complications of Blood Transfusion of Thalassemia Patients	
[17]	P. Paokanta	Bayesian Networks (BNs), Multinomial Logistic Regression, MLP, KNN, and Naïve	classification performance of Machine Learning Techniques for screening β -Thalassemia	Bayesian Networks (BNs) 85.83%, KNN 88.98%, MLP 87.40%, Multinomial Logistic Regression

		Bayes		84.25% and Naïve Bayes 84.25%
[18]	Sacco		A New Approaches for Classifications of Beta Thalassemia.	
[19]	Jahangiri, Mina		Decision-tree-based methods for differential diagnosis of β -thalassemia trait from iron deficiency of anemia.	

2.4 Bangladesh perspective

Our research work is prominently constructing a model analyzing data and applying machine learning algorithms. Our proposed model can predict the thread of Thalassemia disease. This prediction will definitely have a remarkable and important impact on society. People belong to all ages especially adults can be aware of this disease. According to our country's prospective health administration and expenditure are not economic. Patients willing to visit doctors not only suffer due to heavy transportation jams and cost but also harassed by the doctors. The availability of doctors and the hiked visiting gives pain to the pockets of the patients; basically this restrains many lower middle class people visiting the doctors due to the heavy visiting. The doctors can't predict the disease without having a blood test hence the pathological labs dominate this sector with their costly operations. So, our project in this regard tries to ease the difficulties faced by the patients in between doctors and pathological test. This project will help patient's self-assessments of their reports before visiting doctors. Hence it will be economically beneficial for those people who are earning their bread and soul with their tremendous hardship.

2.5 Challenges

We have faced a lot of difficulties while executing this project, prominently in the time of data collection. As we have taken the maximum number of data from an online source the main reason behind it was the lack of availability of pathological real time data. Collecting data practically is too much tough as the pathological labs were not willing to provide the actual authenticated data due to the privacy issues. This is why we have taken the maximum number of data from open online source and used SMOTE technique to increase the number of data and balance them.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The main intention and motive of this research is to develop a model for predicting the Thalassemia positive or negative person. The Prediction Model is made based on medical report of complete blood cell (CBC) and some other relevant info. To build and set this model, we have applied several ML(machine-learning) approaches and algorithms. We have applied KNN, LOGISTIC REGRESSION, SVM, NAÏVE BAYES, DECISION TREE, XGBOOST, ADA BOOST classifier, RANDOM FOREST, MLP, GRADIENT BOOSTING classifier in this research. Algorithms used in the model for classification purposes. 11 relevant factors are utilized that were very closely connected. We analyzed some of the features that were responsible for the outcome. We processed our overall dataset then calculated the accuracy, sensitivity, specificity, precision, recall, F1 score, and roc-curve of each algorithm to select the authentic appropriate performance of algorithm for the model. We found ADA BOOST had the best accuracy and suitable for our proposed model.

3.2 Procedures of Data Collection

The data set is a group or collection of required and relevant coordinates. First of all, we have tried to find out Thalassemia patients in front of our home, school, university and society and in different places. Data collection process was very hard as required data's are not shared by everyone. We have gathered just a few information from using a google form. Again we used datasets from online dataset. Finally, we were capable to gather data of 368 people based on 11 factors. There are 71 Thalassemia positive info and 297 healthy people's info we have. For balancing our dataset, we used SMOTE technique. We gathered all of our data depending on the following factors:

- Age
- Gender

- RBC
- MCV
- PCV
- MCH
- MCHC
- RDW
- PLT
- TLC
- HGB

A1		Age											
	A	B	C	D	E	F	G	H	I	J	K	L	
1	Age	Sex	RBC	PCV	MCV	MCH	MCHC	RDW	TLC	PLT /mm3	HGB	Thalassemia	
2	28	0	5.66	34	60.1	17	28.2	20	11.1	128.3	9.6	0	
3	41	0	4.78	44.5	93.1	28.9	31	13	7.02	419	13.8	0	
4	40	1	4.65	41.6	89.5	28.8	32.2	13	8.09	325	13.4	0	
5	76	0	4.24	36.7	86.6	26.7	30.8	14.9	13.41	264	11.3	0	
6	20	1	4.14	36.9	89.1	27.8	31.2	13.2	4.75	196	11.5	0	
7	24	0	4.29	40.1	93.5	29.6	31.7	14.5	13.96	233	12.7	0	
8	28	1	4.98	42.3	84.9	24.9	29.3	16.2	9.33	213	12.4	0	
9	14	0	4.97	43.8	88.1	28	31.7	15.2	3.92	229	13.9	0	
10	16	0	4.16	38.7	93	28.8	31	17.9	5.77	211	12	0	
11	62	0	5.25	45.6	86.9	25.3	29.2	15.6	10.68	151	13.3	0	
12	42	0	2.17	28.3	93.5	28.1	30	24.6	3.46	92	6.1	1	
13	28	0	4.81	44.4	92.3	27.9	30.2	14.3	6.22	150	13.4	0	
14	59	0	3.41	32.9	96.5	29.9	31	16.8	6.62	132	10.2	0	
15	28	1	2.26	26.9	119	41.2	34.6	15.6	5.27	222	9.3	0	
16	60	0	4.52	38.5	87.4	26.5	30.4	14.7	10.52	589	12	0	
17	22	0	5.17	44.5	86.1	27.7	32.1	13.2	10.7	268	14.3	0	
18	64	0	4.6	41.4	90	28.5	31.6	14.4	9.67	150	13.1	0	
19	78	0	4.24	36.7	86.6	26.7	30.8	14.9	13.41	264	11.3	0	
20	57	0	3.01	31.2	103.7	29.9	28.8	25	2.88	400	9	1	
21	51	0	6.6	49.3	88	27.9	31.6	13.3	9.55	154	15.6	0	
22	50	1	5.46	40.5	74.2	23.3	31.4	17.2	12.05	350	12.7	0	
23	77	1	4.36	32.2	73.9	21.3	28.9	16	10.8	438	9.3	1	
24	40	1	4.56	35.9	78.7	23.5	29.8	15.8	6.17	233	10.7	0	

Figure 3.2.1 Collected dataset demo table for predicting thalassemia

3.3 Application of ‘SMOTE’ Technique:

Imbalanced classification needs developing predictive models on classification imbalanced dataset. Imbalanced dataset effects on performance of the predictive model. Oversampling of minority can be done by one significant approach. The simplest method is making duplicates examples minority class. This duplicating method don't create any new information that can create extra effect to the model. Instead, new examples can be synthesized from that already

existing one's. This can be data augmentation for the minority class and is referred to as the SMOTE or Synthetic Minority Oversampling Technique.

We applied this method on our imbalanced dataset so that it can be used as a balanced dataset. Before applying this smote technique our dataset contained 368 data after applying this it increased and became almost 594.

3.4 Research Subject and Instrumentation

ML algorithms and methods, data mining and deep learning are very renown, famous for any prediction and detection at present. We have implemented various algorithms to our collected dataset to see which algorithms will forecast best for our model. We have used several ML algorithms; they are KNN, LOGISTIC REGRESSION, SVM, NAÏVE BAYES, DECISION TREE, RANDOM FOREST, MLP, XGBOOST, ADA BOOSTING classifier and GRADIENT BOOSTING classifier. We used 'Python', the popular programming language and 'GOOGLE COLAB' and 'Microsoft Excel' as our dataset in our research work.

3.4.1 Proposed Methodology

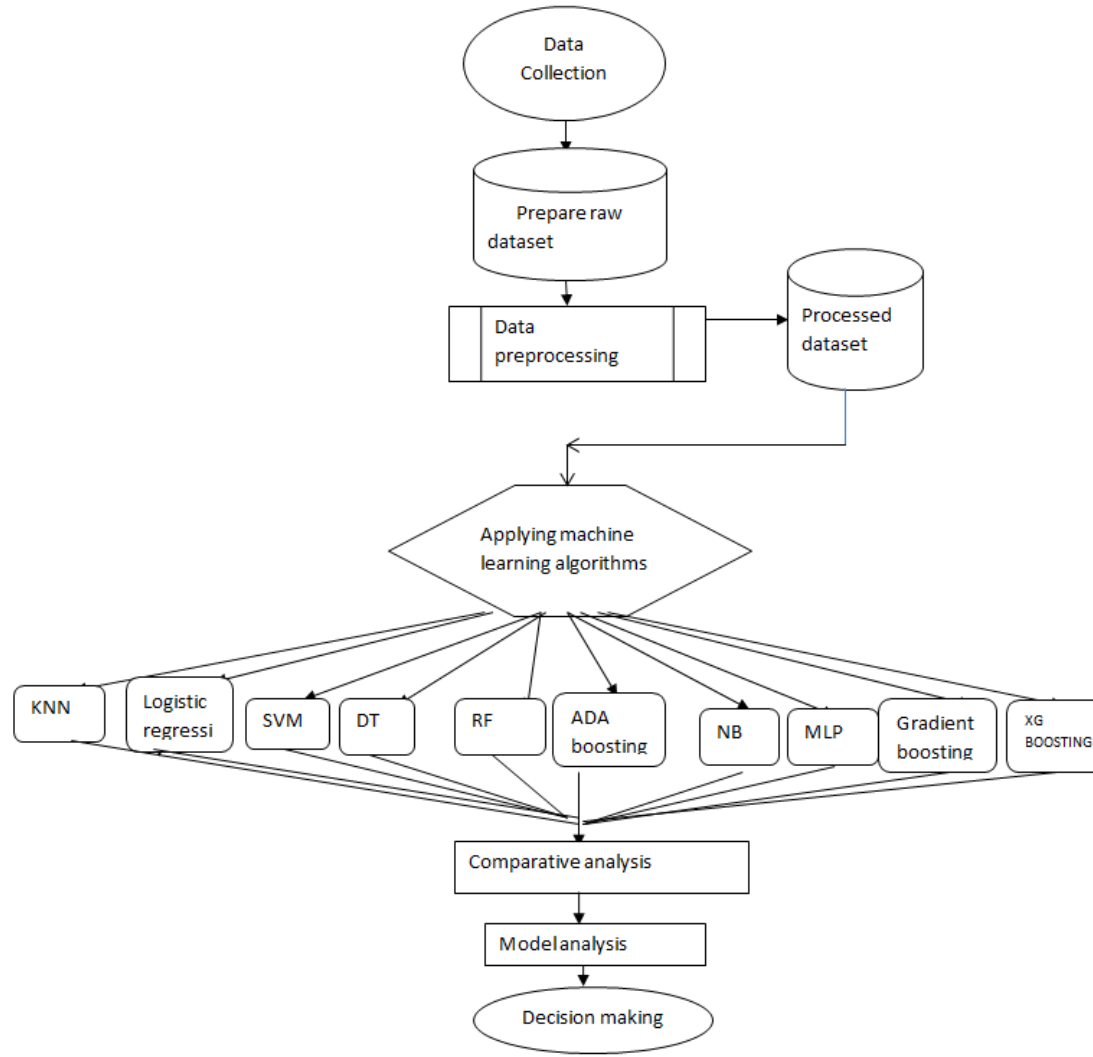


Figure 3.1: Steps of our proposed methodology

3.4.2 Data Preprocessing

After storing and gathering all the data, we found that dataset is not properly appropriate for our algorithm. There we faced problem with some data that missed, categorical , numerical and text data. Our dataset needed to be preprocessed, and then data will be adjustable to our desired ML algorithms. Data processing is the approach to transform data into a suitable format after collection of data. Processing in a specific format aids one to make result easier.

Data preprocessing technique is described below in Figure 3.2

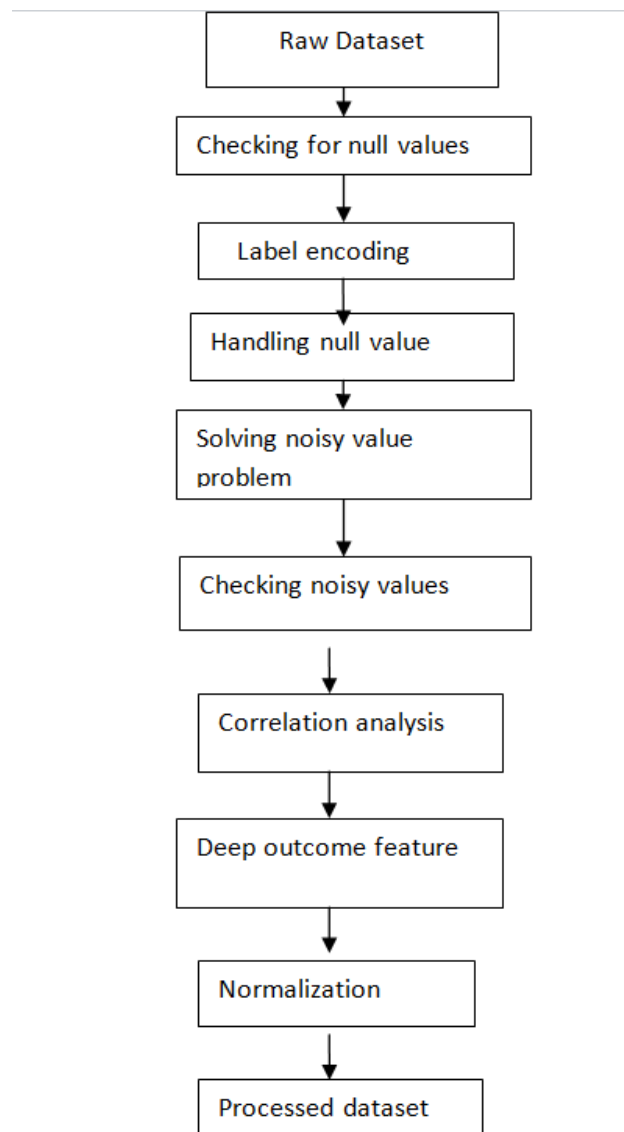


Figure 3.2: Steps of data preprocessing

Firstly, we initialized the process of data cleaning. We checked dataset if there are any duplicate values or not. Then we noticed for a null value or not in the dataset. By encoding the level which will convert the all text to numerical. We solved the issue of missing values dropping incomplete data as our dataset is not larger one. After that we have carefully checked if there is a noisy value in the data set. We have analyzed the correlation matrix using data integration process. The ratio of connections

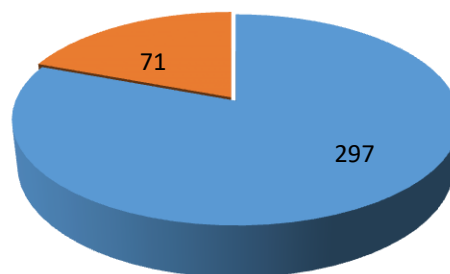
of details shown by this matrix. We have analyzed some scatter matrices, parallel Coordinates, on features like (Age ,RBC,PCV,MCV,MCH,TLC,RDW,MCHC,HGB,PLT ETC).Even we have analyzed some heat map on feature (Age, Hgb, Pcv, Sex, Thalassemia).Our data transformation was prepared using normalization technique. Thus, After all this type of processes finally we got the processed data set in our hands. This whole process of data processing was done using the “Google Colab”.

3.5 Statistical Analysis

We were able to gather a dataset of 368 data there 297 people were Thalassemia positive and 71 people were negative. For balancing this dataset we have used oversampling method called “SMOTE Technique”.

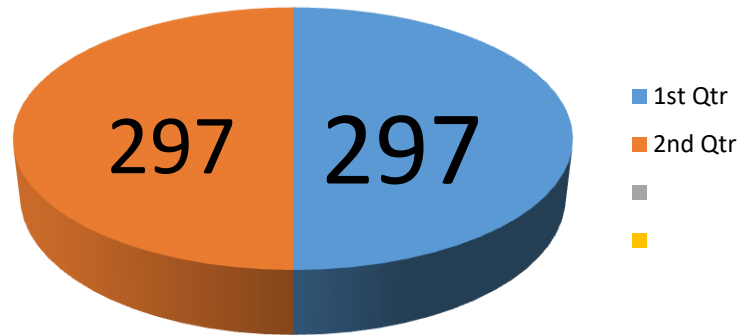
After that 297 people were Thalassemia positive and 297 people were Thalassemia negative. Figure 3.3 and 3.4 depicts that in our dataset how many Thalassemia positive and negative people were. We had made our model which was based on data from 297 Thalassemia positive and 297 Thalassemia negative person.

Thalassemia prediction



3.5.1 Before using smote technique Thalassemia positive or negative cases (26% negative and 84 % positive)

Thalassemia prediction



3.5.2 After using smote technique Thalassemia positive or negative cases (50% positive and 50% negative).

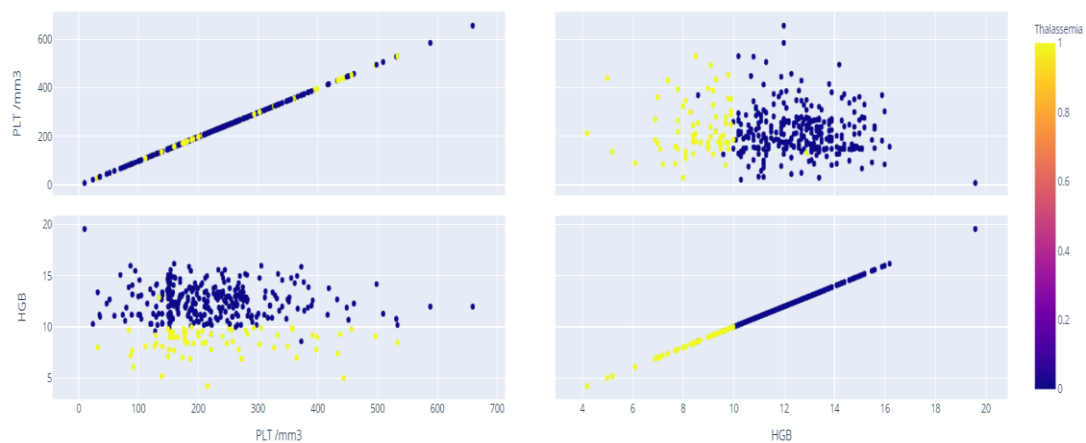


Figure 3.5.3 PLT and HGB case

Figure 3.5.3 shows that information from people of some feature like PLT, HGB. These points are clearly separable for classification.

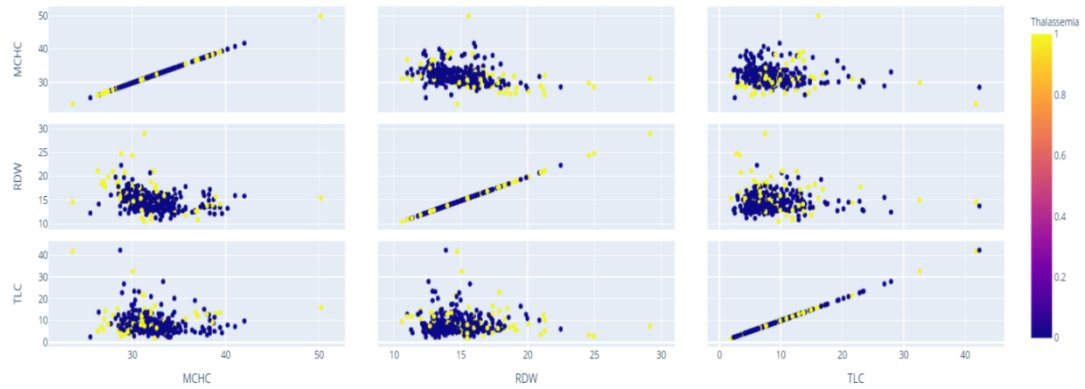


Figure 3.5.4 TLC, MCHC, RDW

Figure 3.5.4 shows that information from people of some feature like TLC, RDW, and MCHC. Again these points are clearly separable for classification.

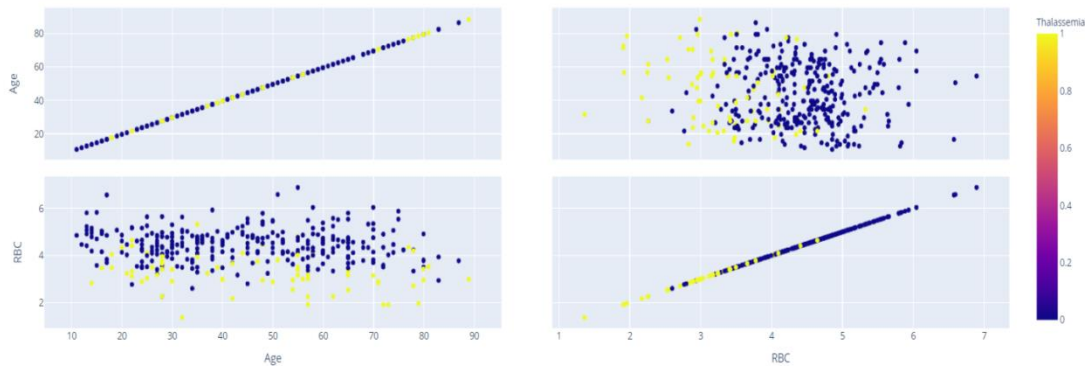


Figure 3.5.5 RBC, AGE

Figure 3.5.5 shows that information from people of some feature like RBC, Age. Points can be used for classification also.

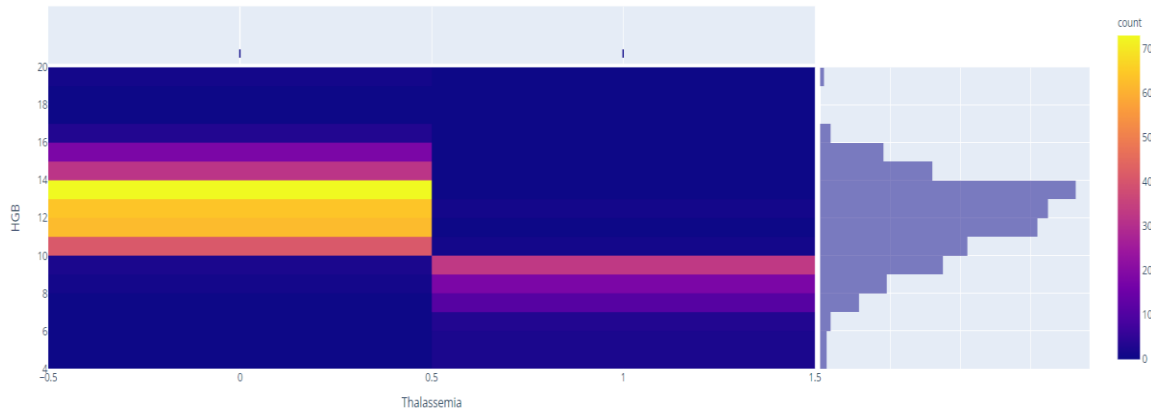


Figure 3.5.6 HGB vs Thalassemia

Figure 3.5.6 shows a histogram graph analysis that information from people about Thalassemia positivity on HGB.

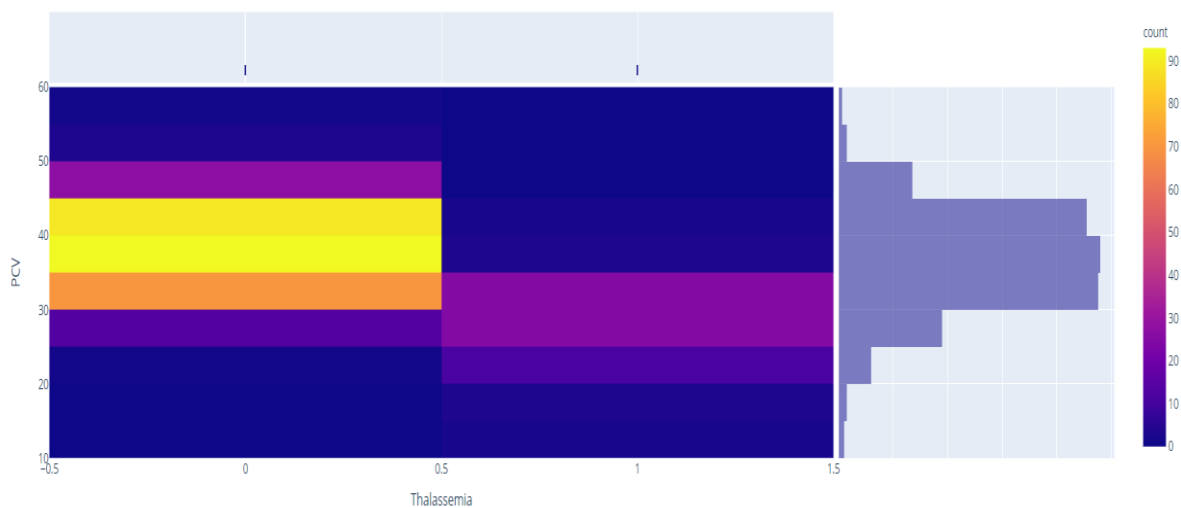


Figure 3.5.7 PCV vs Thalassemia

Figure 3.5.7 shows a histogram graph analysis that information from people about Thalassemia positivity on PCV.

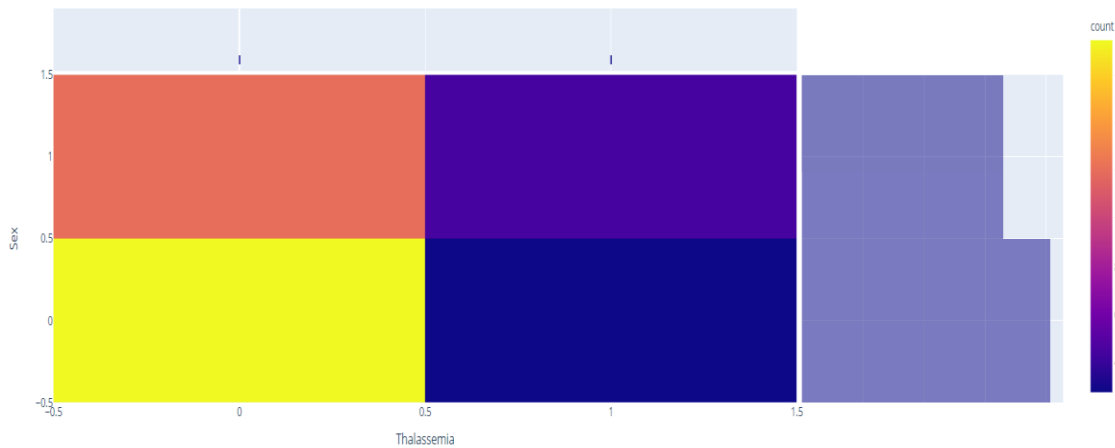


Figure 3.5.8 SEX vs Thalassemia

Figure 3.5.8 shows a histogram graph analysis that information from people about Thalassemia positivity on SEX.

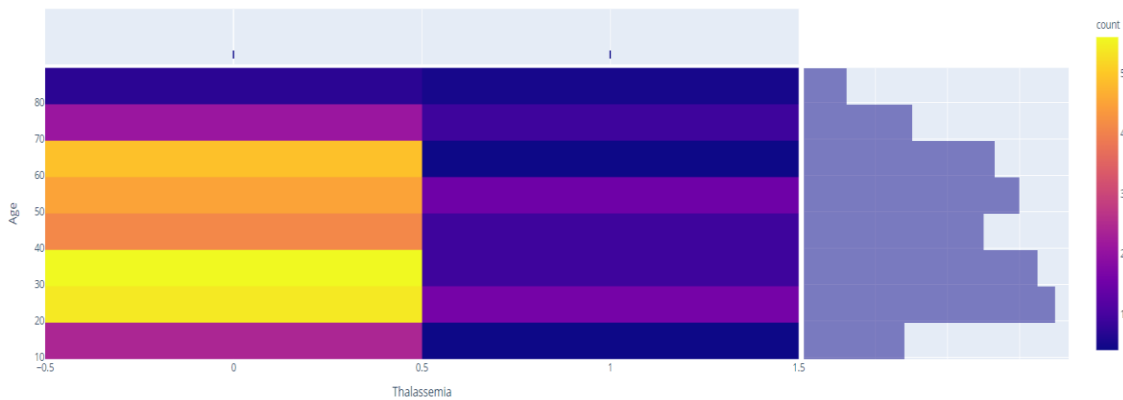


Figure 3.5.9 AGE vs Thalassemia

Figure 3.5.9 shows a histogram graph analysis that information from people about Thalassemia positivity on AGE.

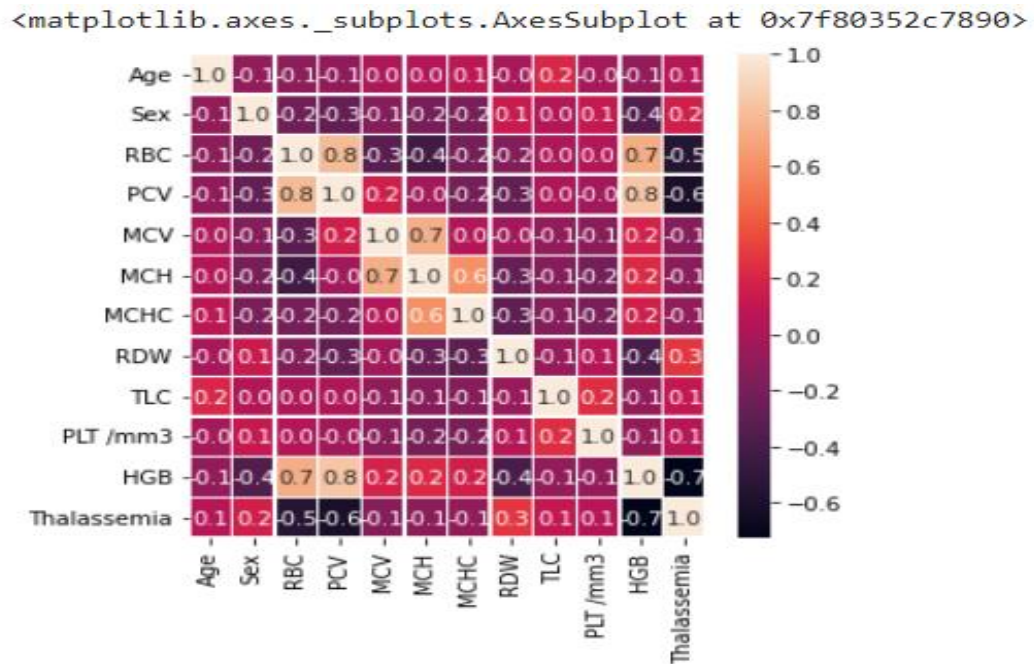


Figure 3.5.10 Correlation Matrix.

Figure 3.5.10 depicts the correlation between the features. This correlation matrix describes the features connectivity to others feature.

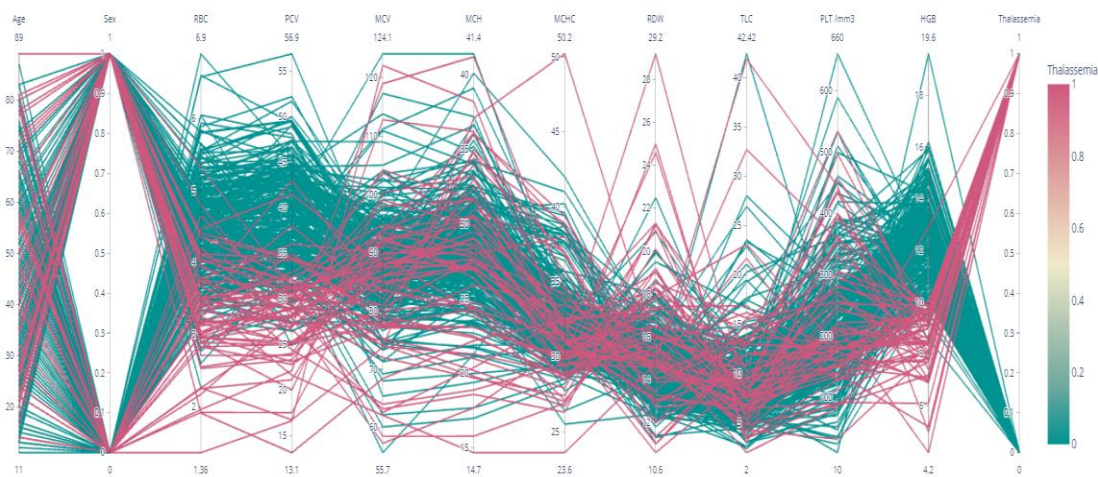


Figure 3.13 Parallel Co-ordinates.

Figure 3.13 depicts the overall relations between the features. It gives us the overall result about our used dataset .Anyone can get enough explanation looking this at a glance.

CHAPTER 4

PERFORMANCE OF THE PROPOSED MODEL

4.1 Introduction

We reviewed the dataset and its processing methods in the preceding segment. The processed data is employed in several algorithms, which will be detailed in this chapter. The techniques KNN, LOGISTIC REGRESSION, SVM, NAVE BAYES, DECISION TREE, xgboost, RANDOM FOREST, MLP, ADA BOOSTING predictor, and GRADIENT BOOSTING classification algorithm are all utilized, and the results are analyzed to see which algorithm provides the highest accuracy. There are a few procedures required to compute the accuracy. We accumulated 375 data points from individuals that were either Thalassemia present or absent, with 70% being utilized as training phase and 30% for use as test data. Our dataset is called 'Thalassemia Dataset.' We have applied SMOTE technique also here for balancing our dataset.

4.2 Results & Analysis of Experiments

We analyzed the accuracy, confusion matrix, precision, recall, F1 score, sensitivity, and specificity of 10 machine-learning techniques.

4.2.1 Experimentation

To begin, we employ K fold cross-validation. Cross-validation is a method that resembles for evaluation of ML models on a small sampled data. The procedures includes only one parameter, k, which means or specifies the num of groups into which a given data sample should be in division. It's a popular technique since it's straightforward to grasp and produces a les biased or optimistic estimation of model competence than other aproaches, such as a simple train/test division. kNN has 93 percent accuracy, SVM gives 98 percent accuracy, logistic regression has

98 percent accuracy, Naïve Bayes has achieved 93% accuracy, the Random forest has 97 percent accuracy, decision tree has 97 percent accuracy, MLP has 97 percent accuracy, Stochastic Gradient boosting classifier has 97 percent accuracy, and XGBoost has 97 percent accuracy and finally ADA boosting classifier has 100 percent accuracy.

4.2.2 Analytical Methodology

We calculated the sensitivity, specificity, precision, recall, f-score, roc-curve, and confusion matrix of each algorithm in addition to their accuracy. Any model selection must include an evaluation of that model. Certain classifiers must be measured. For better measurement, classifications are measured using the test dataset. The real positive rate is defined as Sensitivity. That is, sensitivity is defined as the ratio of successfully identified positive tuples to the total number of instances item sets. The genuine negative rate is known as specificity. That seems to be, specificity is the ratio of accurately determining negative tuples to the total number of adverse tuples. Precision refers to the evaluation of perfection. It's the proportion of true positive to expect positive.

The term "recall" refers to the reliability of accuracy. The ratio of true positive value to true positive value is known as the true positive value-to-true positive value ratio.

The harmonic mean of recall and precision is measured by the F1 score. It considers both false detection values for calculation.

The graphical evaluation of categorization models using receiver operating characteristics (roc) curves is particularly beneficial. True positive and false-positive rates are used to create the ROC curve. The random guessing is symbolized by the diagonal line. A model's curve is similar to guessing at random, which is a less accurate model. As a result, the curve for an accurate model willn't close from the guessing line. Our utilizing algorithms' ROC curves are shown in the figures below.

➡ No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.993

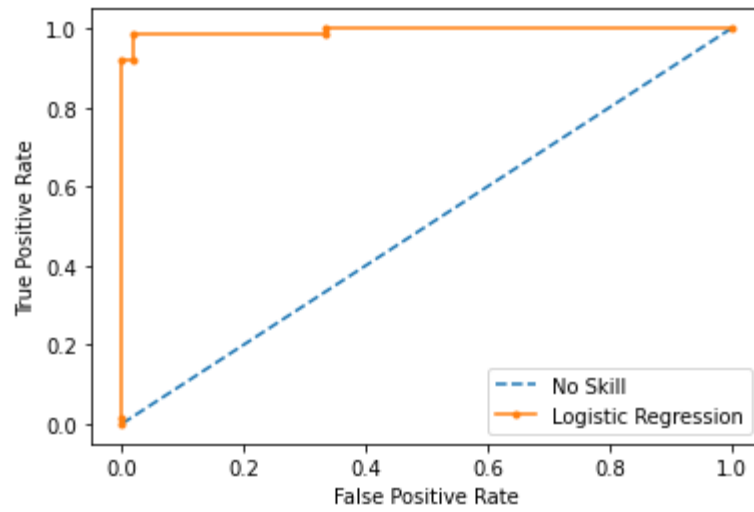


Figure 4.1: ROC curve of Logistic Regression Algorithm

➡ No Skill: ROC AUC=0.500
Logistic: ROC AUC=0.974

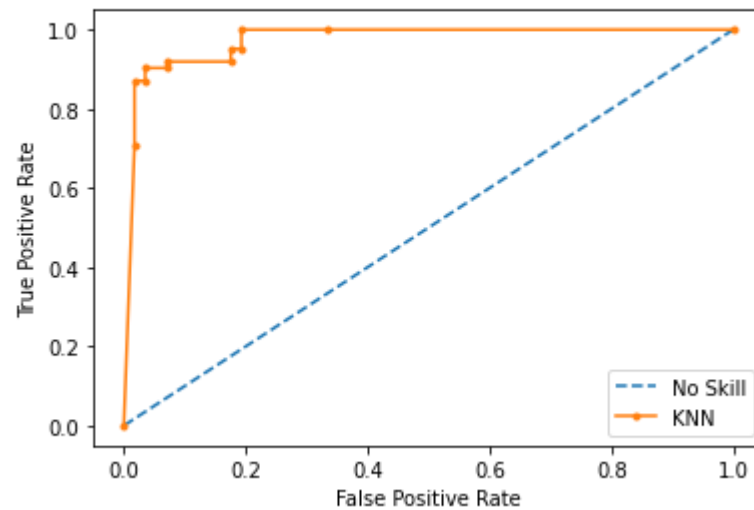


Figure 4.2: ROC curve of kNN Algorithm

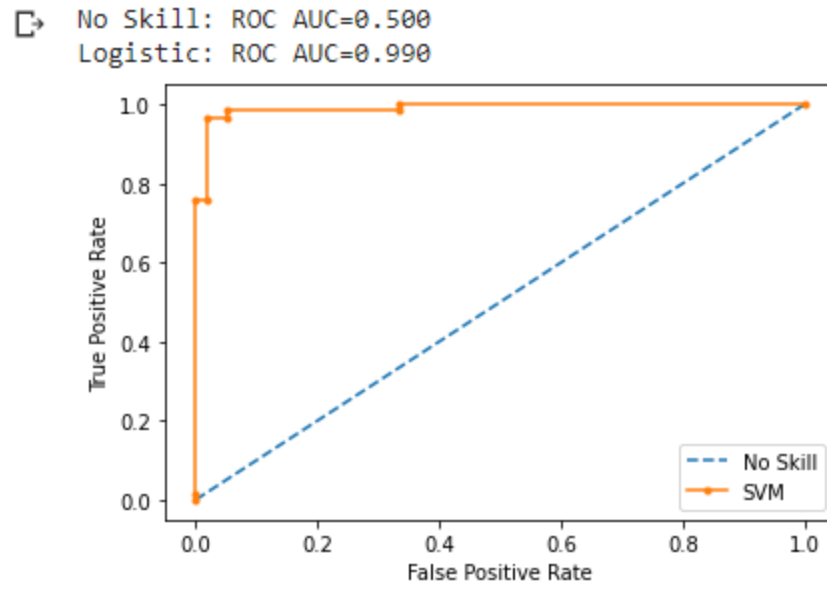


Figure 4.3: ROC curve of SVM Algorithm

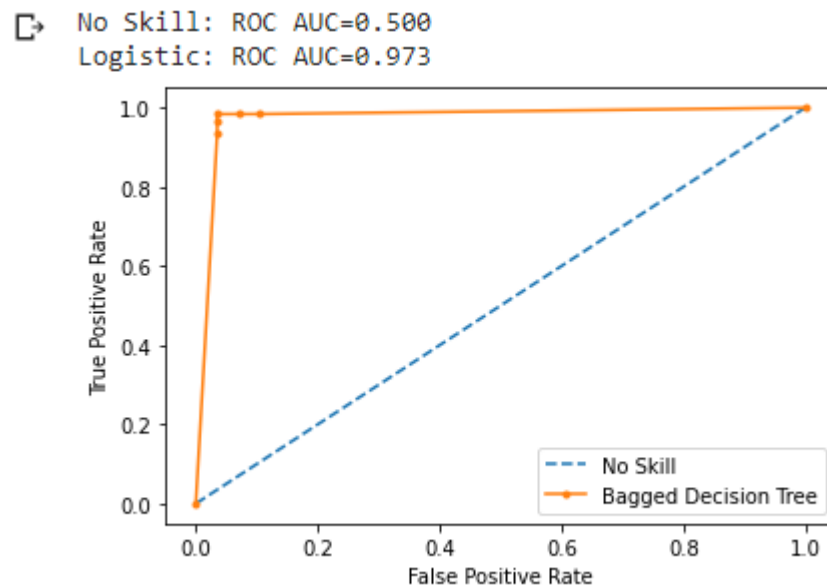


Figure 4.4: ROC curve of Bagged Decision Tree Algorithm

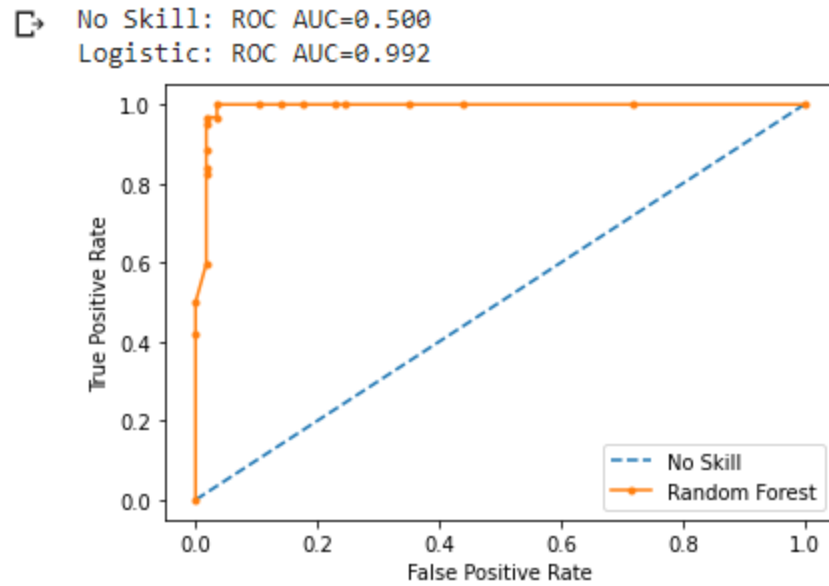


Figure 4.5: ROC curve of Random Forest Algorithm

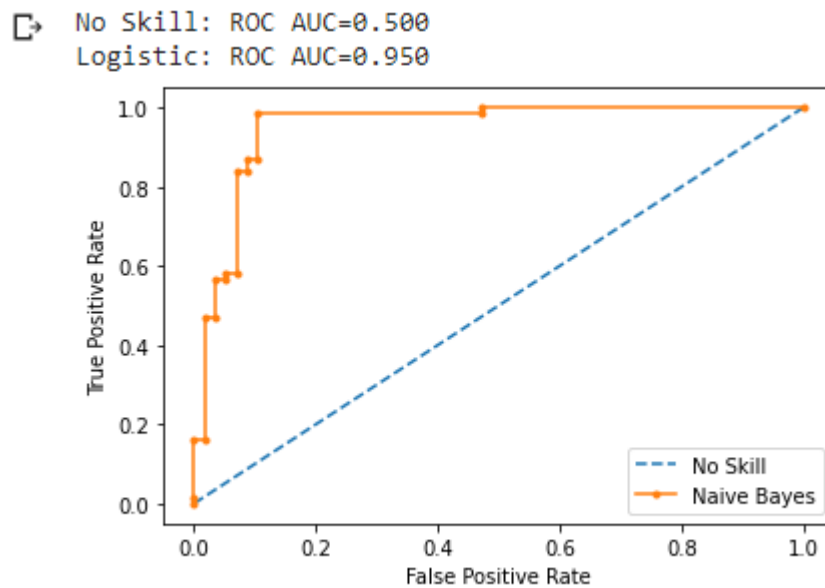


Figure 4.6: ROC curve of Naïve Bayes Algorithm

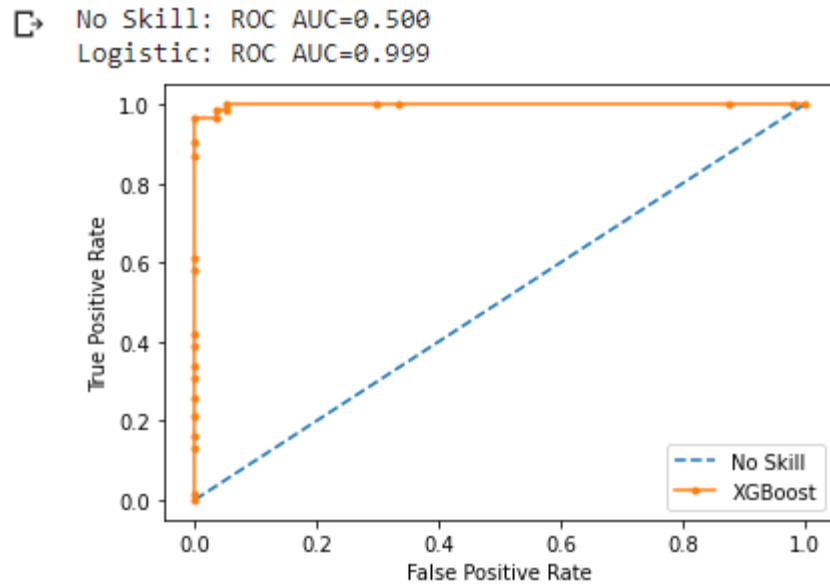


Figure 4.7: ROC curve of XGBoost Algorithm

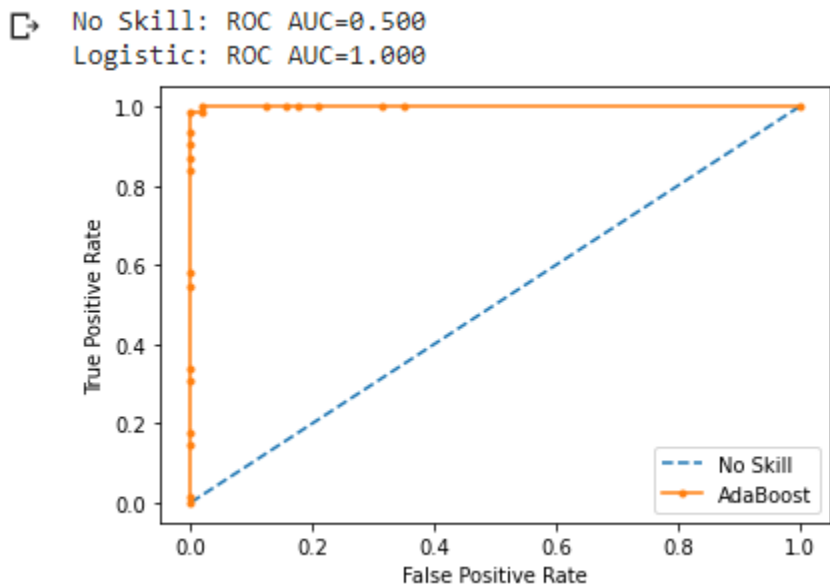


Figure 4.8: ROC curve of AdaBoost Algorithm

No Skill: ROC AUC=0.500
 Logistic: ROC AUC=0.978

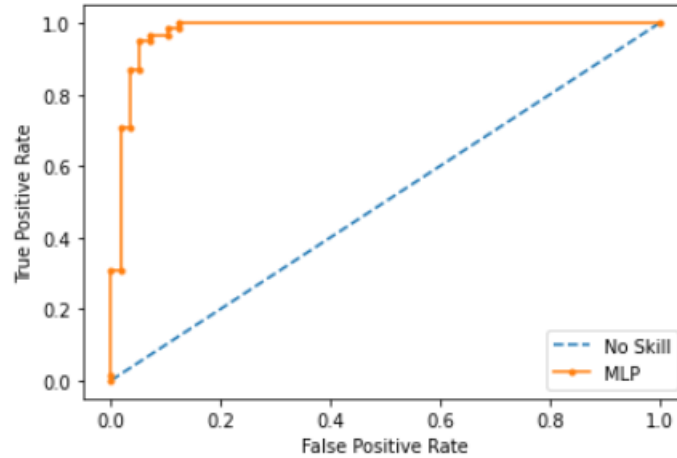


Figure 4.9: ROC curve of MLP Algorithm

➡ No Skill: ROC AUC=0.500
 Logistic: ROC AUC=0.989

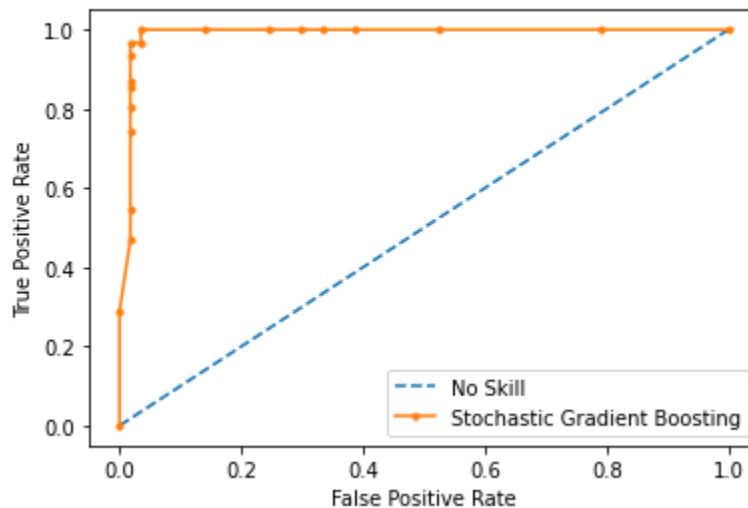


Figure 4.10: ROC curve of Stochastic Gradient Boosting Algorithm

Here in the Logistic Regression ROC AUC curve, we can see the comparison between the average accuracy and the accuracy of ROC AUC curve. In the LR method the accuracy, precision recall and F1 score all are 97%. But the ROC AUC curve shows the accuracy about

99.3%. Where in kNN algorithm the accuracy comes 87%, precision comes 88%, recall comes 87% and F1 score comes 87% but the ROC AUC curve gives the accuracy around 97.4%. Again in SVM the all the diameters shows that the percentage is about 96. Whereas it's ROC AUC curve says it's accuracy is 99%. Bagged Decision Tree comes next and shows that other that precision which is 98% rest of the parameters is 97 percent, in which the ROC AUC curve gives 97.3% accuracy. In Random Forest algorithm the accuracy, recall and F1 score all are 97% respectively but the precision is 98%, where its ROC AUC curve defines the accuracy by 99.2%. Naïve Bayes algorithm denotes 93 percent of accuracy for recall, F1 score and the average accuracy but for precision it is 94% and in the ROC AUC curve it employ 95% accuracy. Where for XGBoost technique the diameters without precision gives 97% accuracy and for precision it is 98%, ROC AUC defines 99.9% of average accuracy. In AdaBoost method, the results shows its best performance while calculating the accuracy, precision, F1 score and recall values and it is about 99% and the good thing is that the ROC AUC curve of it also shows the highest accuracy about 100%. Which means the prediction using AdaBoost technique was worth in this case. The MLP And stochastic the contradictory results for the diameter measurements MLP gives 97 percent for all but for the Stochastic Gradient Boosting Method those was 96% rather than precision which was 97% and the final ROC curve allows that accuracy about 98.9%.

Hence, according to the above discussion we can conclude as the algorithm results were one on the top of other. Among the techniques the AdaBoost algorithm gives the highest accuracy for ROC AUC curve which is the main concern of prediction Thalassemia Disease. The other algorithms could beat the rest but for the value of their curve they couldn't. If once accuracy beat another its curve didn't give the best accuracy this type of situation happened. But in the case of AdaBoost algorithm the Diameters accuracy plus its Curve's accuracy was the highest.

Most significant performances measuring approaches for ML classification is the confusion matrix. It will run the classification models against the test data and output the tp, tn, fp, and fn values in a statistical manner. The Confusion Matrix is critical for evaluating any classifier's performance. Here we also have done confusion metrics to evaluate our analysis more clearly.

Table 4.2 depicts the confusion matrix of all algorithms used in our model.

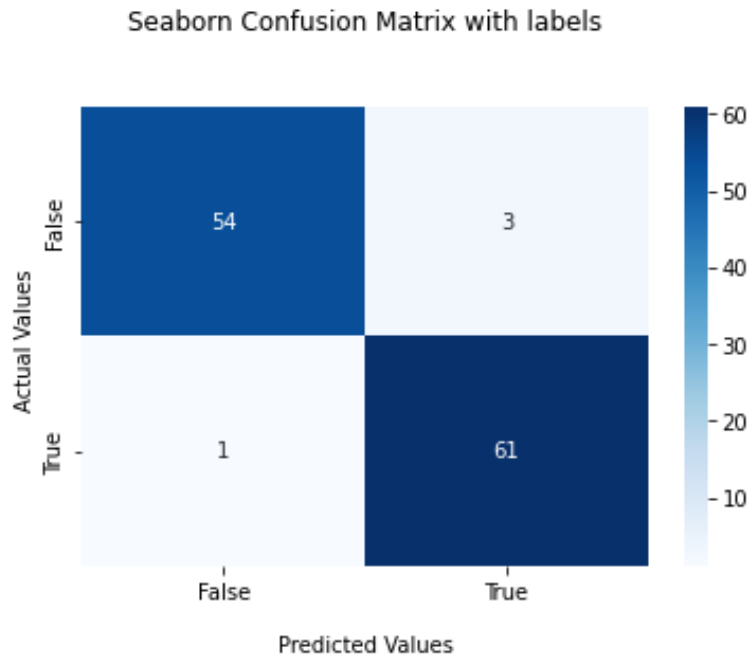


Fig 4.11: Confusion Matrix of Logistic Regression Algorithm

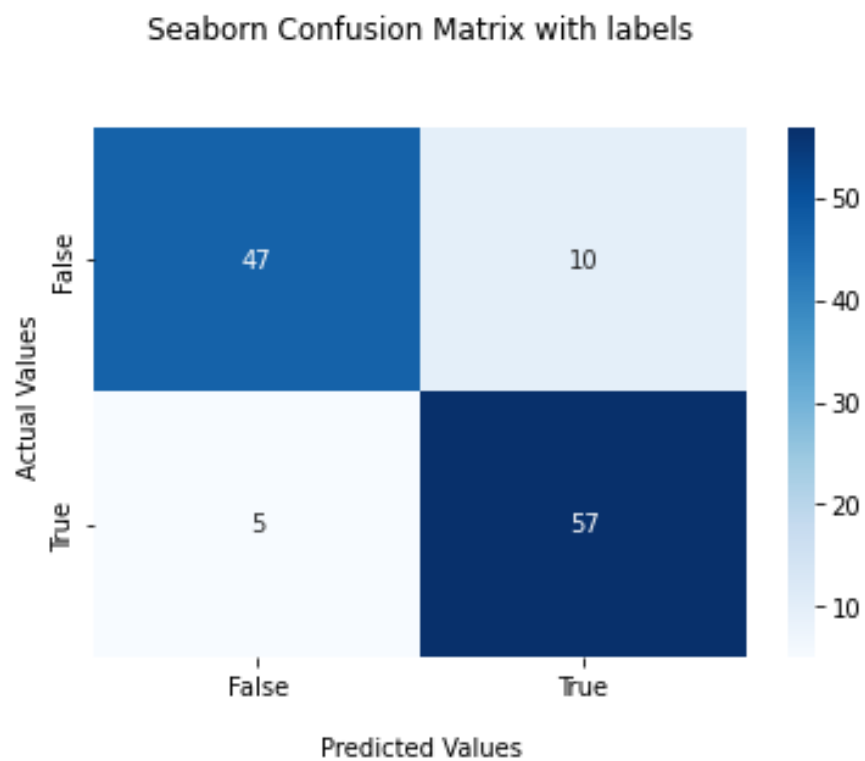


Fig 4.12: Confusion Matrix of K-Nearest Neighbor Algorithm

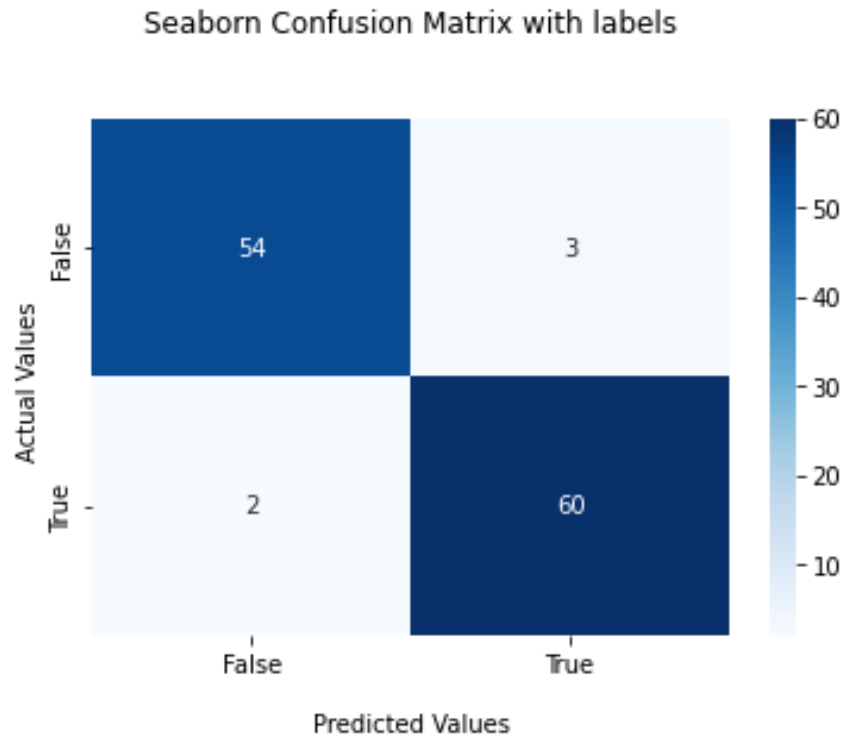


Fig 4.13: Confusion Matrix of Support Vector Machine Algorithm

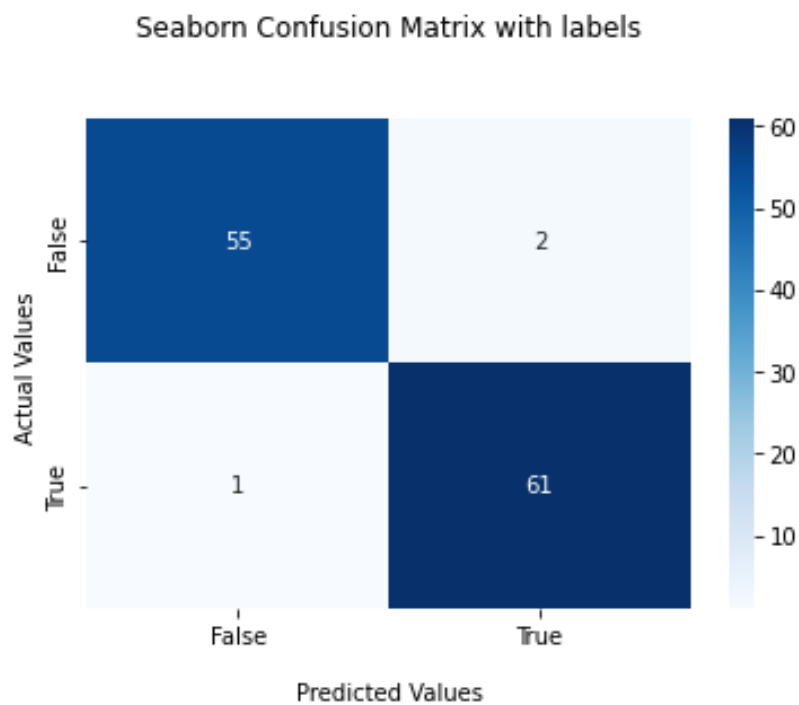


Fig 4.14: Confusion Matrix of Bagged Decision Tree Algorithm

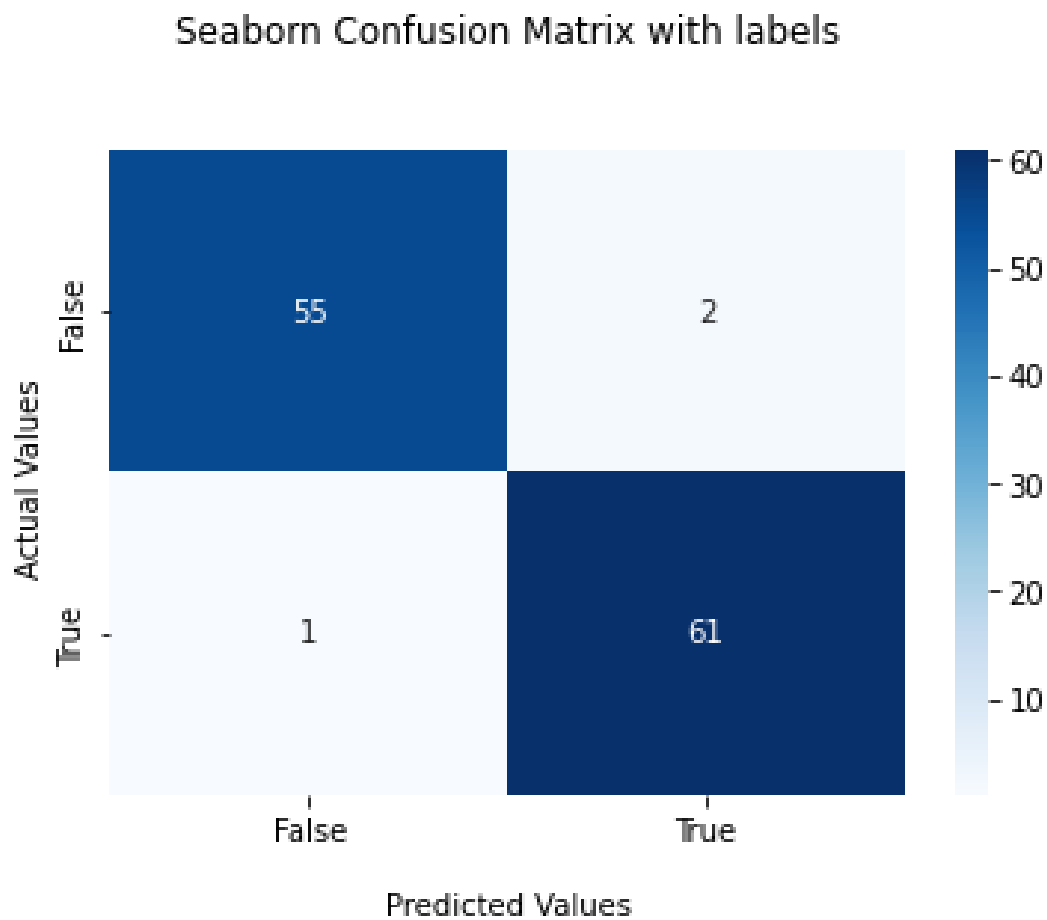


Fig 4.15: Confusion Matrix of Random Forest Algorithm

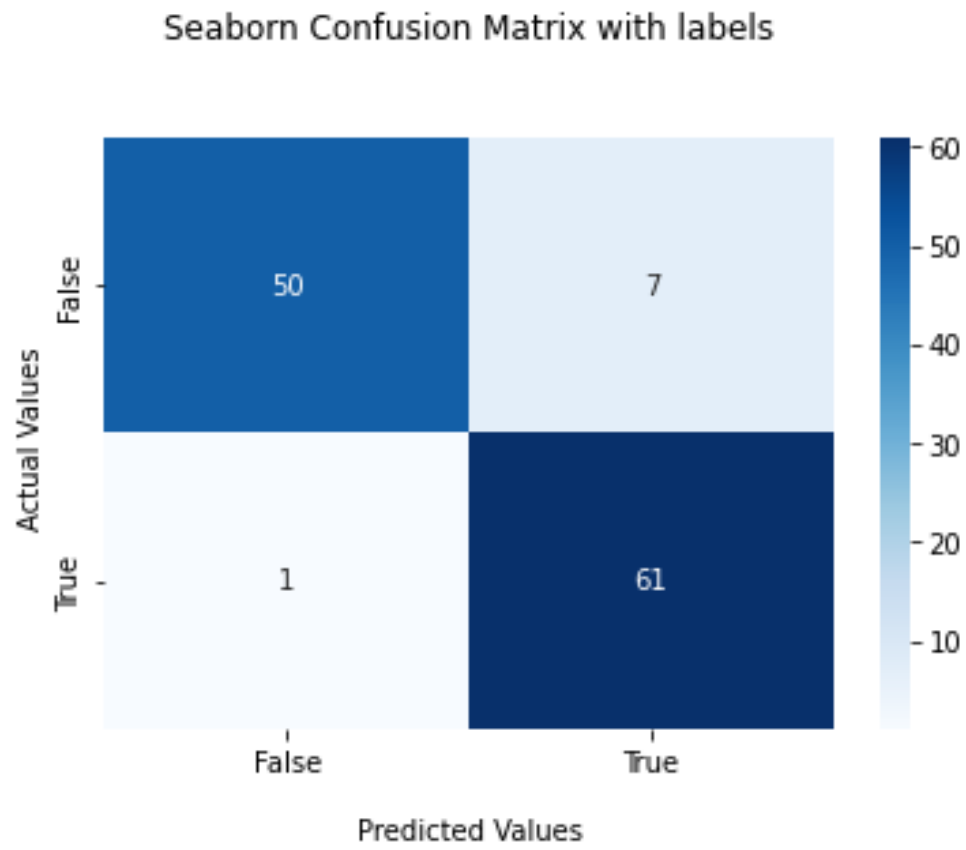


Fig 4.16: Confusion Matrix of Naïve Bayes Algorithm

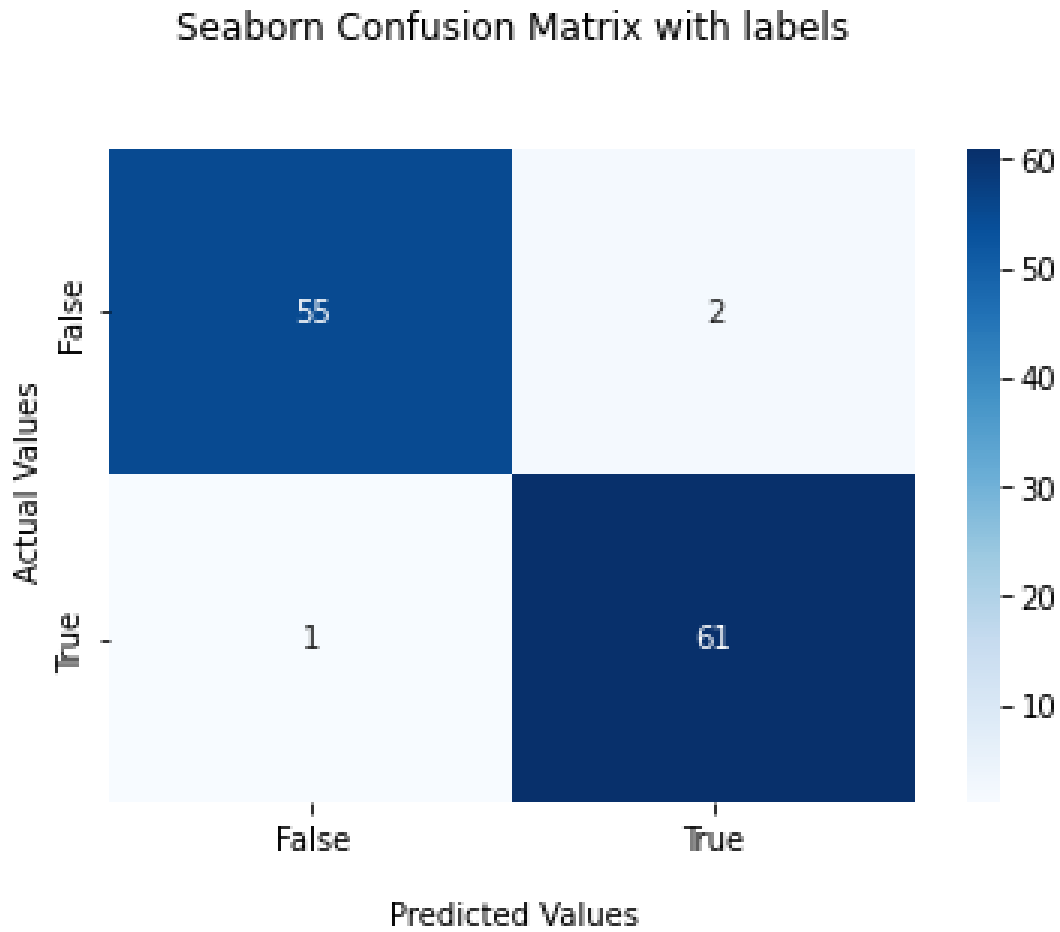


Fig 4.17: Confusion Matrix of XGBoost Algorithm

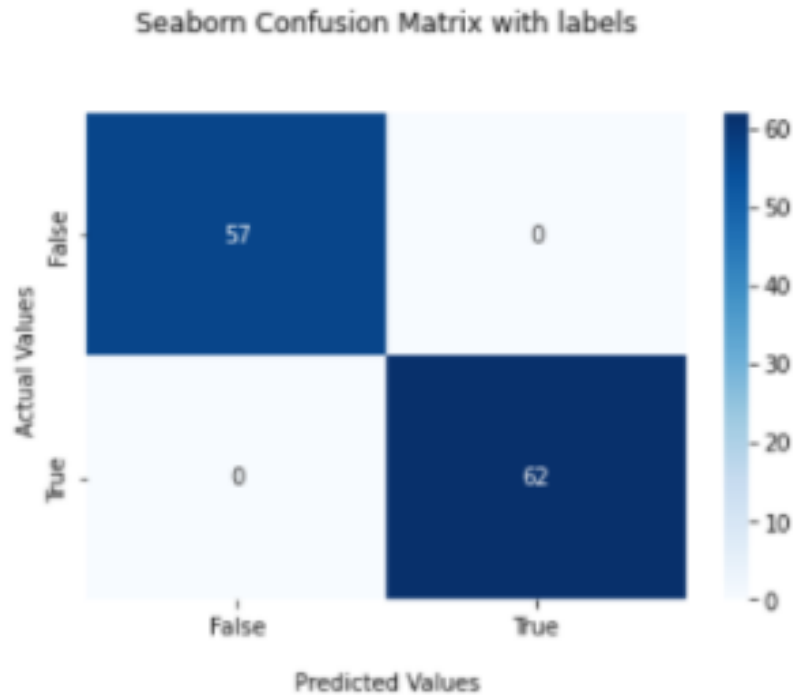


Fig 4.18: Confusion Matrix of AdaBoost Algorithm

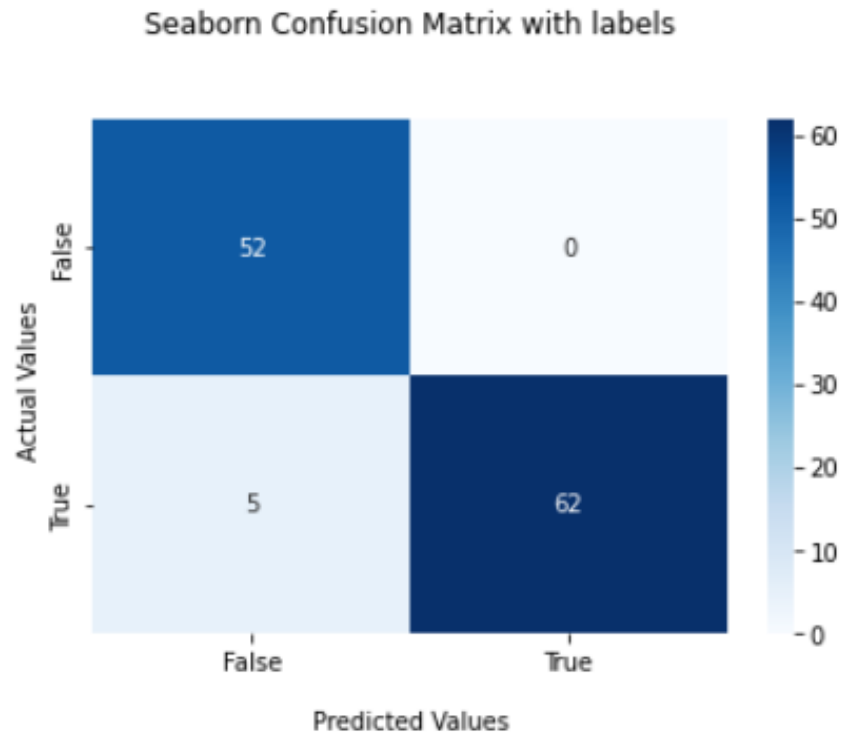


Fig 4.19: Confusion Matrix of Multilayer Perceptron Algorithm

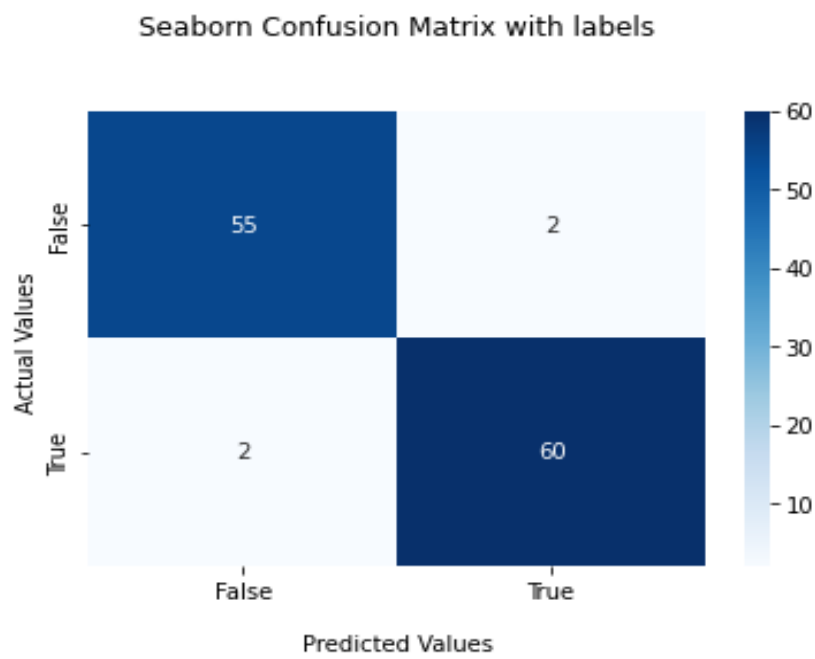


Fig 4.20: Confusion Matrix of Stochastic Gradient Boosting Algorithm

Confusion matrix denotes the predicted value versus actual value comparison. True positive (TP) denotes in the case of Thalassemia that the patient doesn't have Thalassemia and the machine predicted negative for that. Then False Positive (FP) denotes that the patient doesn't have Thalassemia but the prediction of the machine is positive. In case of having Thalassemia in a person where the algorithmic prediction is negative then the case is called False Negative (FN). Lastly If a patient found Thalassemia positive and the method also says Yes that person has the disease then the term is called True Positive (TP). Here the Confusion matrix for the Ten algorithms we have used shows that, In Logistic Regression the True negative value is 54, True positive is 61, False positive value is 3 and False negative is 1. In KNN the True negative value is 47, True positive is 57, False positive value is 10 and False negative is 5. Confusion matrix of SVM shows that, True negative value, True positive, False positive value and False negative value are 54, 60, 3, 2 respectively. In DT algorithm, True negative value is 55, True positive is 61, False positive value is 2 and False negative is 1. 55, 61, 2, 1 are the True negative value, True positive, False positive value and False negative values of RF respectively. The True negative value, True positive, False positive value and False negative value of NB are 55, 61, 7, 1 respectively. The confusion matrix of XGBoost shows 55, 61, 2, 1 values of True negative value, True positive, False positive value and False negative values respectively. In AdaBoost algorithm the values are 56, 62, 1, 0 of True negative value, True positive, False positive value and False negative values respectively. In MLP the confusion matrix shows that the True negative value is 52, True positive is 62, False positive value is 0 and False negative is 5. And lastly in Stochastic Gradient Boosting algorithm the TN value, TP, FP value and FN values are 55, 60, 2, 2 respectively.

CHAPTER 5

RESULT COMPARISON AND ANALYSIS

This research provides us the highest accuracy compared to other paper. There are some researcher who worked nearly closed to this work. Table 5.1 reveals the classifier performance evaluation and Table 5.2 exhibitions of some Interrelation between our work and some previous work on several insect predictions.

Table 5.1: Classifier Performance Evaluation

Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
LR	97	97	97	97
KNN	87	88	87	87
SVM	96	96	96	96
DT	97	98	97	97
RF	97	98	97	97
NB	93	94	93	93
XGBoost	97	98	97	97
AdaBoost	100	100	100	100
MLP	97	99	99	99
Stochastic Gradient Boosting	96	97	96	96

Table 5.2 Comparative Analysis

Article	Author name	Used Algorithm	Prediction	Accuracy
[1]	This Work	KNN, LR, SVM, DT, NB, RF, XGBoost, AdaBoost, MLP, Stochastic Gradient Boosting	Thalassemia positive or negative	KNN(87%), LR(97%), SVM(96%), DT(97%), NB(93%) RF(97%), XGBoost(97%), AdaBoost(100%), MLP(97%), Stochastic Gradient Boosting(96%)
[2]	Eyad H. Elshami	Decision Tree, Naïve Bayes, Neural Network	Diagnosis of Thalassemia	Decisiontree (93.64%), naive Bayes (93.7%), and Neural Network (95.71%)
[3]	NgoziChidozieEgejuru	WEKA (software decision tree based model), Naïve Bayes, multilayer perceptron	Thalassemia risk prediction	Naïve Bayes (94.12%), multilayer perceptron (100%).
[4]	Yi-Kai Fu	vector	Differentiating	SVM 95%.

		machine learning (SVM)	Thalassemia and Non-Thalassemia Patients	
[5]	FatemehYousefian	KNN, MLP, NN, DT and SVM	Prediction Thalassemia	KNN (93.2%), MLP (99.12%), NB (94.35%), DT (93.64%), NN (95.71%).
[6]	V. Laengsri	k-NN, decision tree (DT), random forest (RF), artificial neural network (ANN) and support vector machine learning (SVM)	discriminating thalassemia trait and iron deficiency anemia	Got High accuracy in (RF, SVM), Poor accuracy in (KNN, DT).
[7]	Roberta Risoluti	TGA/chemometrics method	Update onthalassemiadiagnosis	got full accuracy.
[8]	F R Aszhari	Random forest(RF)	Classification of thalassemia	70-85%
[9]	Alaa S. AlAgha		Identifying β -thalassemia	

[10]	Dr. C.A.D.M.N.C. Kolambage	Random Forest, Artificial Neural Network (ANN)	Predictive Modelling Tool to Accurately Predict Thalassemia Carrier state using Full Blood Count Indices and Hemoglobin Variants.	
[11]	BetülÇil, HakanAyyıldız	Logistic Regression, KNearest Neighbors, Support Vector Machine, Extreme Learning Machine and Regularized Extreme Learning Machine classification	Discrimination of β - Thalassemia and Iron Deficiency Anemia	96.30% accuracy for female, 94.37% for male, and 95.59% in co-evaluation of male and female
[12]	Paokanta, Patcharaporn,	The Multi- Layer Perceptron (MLP). K- Nearest Neighbors (KNN), Naive Bayes,	The Knowledge Discovery of β - Thalassemia	MLP 86.61%, KNN 85.83%, Naive Bayes 85.04%, Bayesian Networks 85.04% and Multinomial Logistic Regression

		Bayesian Networks (BNs) and Multinomial Logistic Regression		82.68%
[13]	P. Paokanta	Bayesian Networks (BNs), Multinomial Logistic Regression, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP) and Naïve Bayes	classification performance of Machine Learning Techniques for screening β -Thalassemia	Bayesian Networks (BNs) 85.83%, Multinomial Logistic Regression 84.25%, K-Nearest Neighbors (KNN) 88.98%, Multi-Layer Perceptron (MLP) 87.40% and Naïve Bayes 84.25%

5.1 Project Interface



The image shows a web form titled "Thalassemia Diagnostic Center". It features a vertical list of input fields for the following parameters: Age, Sex, RBC, PCV, MCV, MCH, MCHC, RDW, TLC, PLT /mm3, and HGB. Below these fields are two buttons: a blue "Submit" button and a blue button with the text "Blood Bank" in red. The entire form is set against a dark blue background.

Fig 5.1: The Interface of Web Project Before Entering Values

Thalassemia Diagnostic Center

28
1
3.68
43
83
34.8
35.7
12
11
177
8.7

Submit

[Blood Bank](#)

Patient found Thalassemia positive!

Fig 5.2: After inserting values of thalassemia positive patient

Thalassemia Diagnostic Center

40
1
4.65
41.6
89.5
28.8
32.2
13.0
8.09
325.0
13.4

Submit

[Blood Bank](#)

Patient found Thalassemia negative!

Fig 5.3: After inserting values of thalassemia negative patient

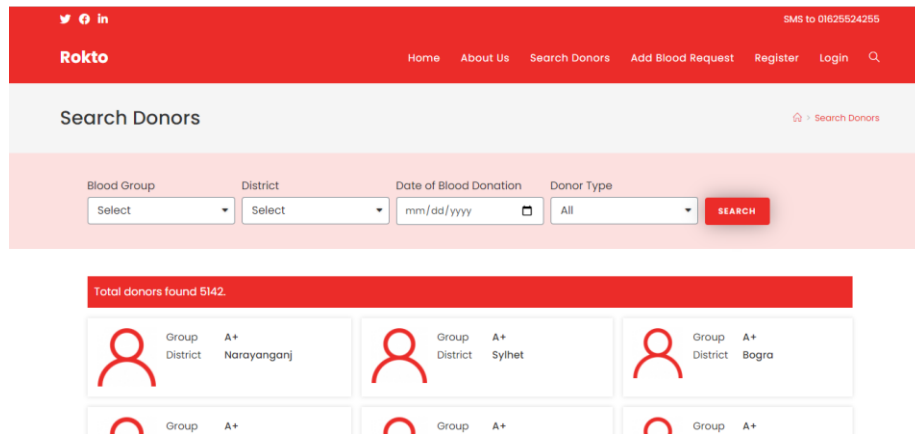


Fig 5.4: Website interface which is linked with our project for finding blood in case of emergency

5.2 Project Architecture

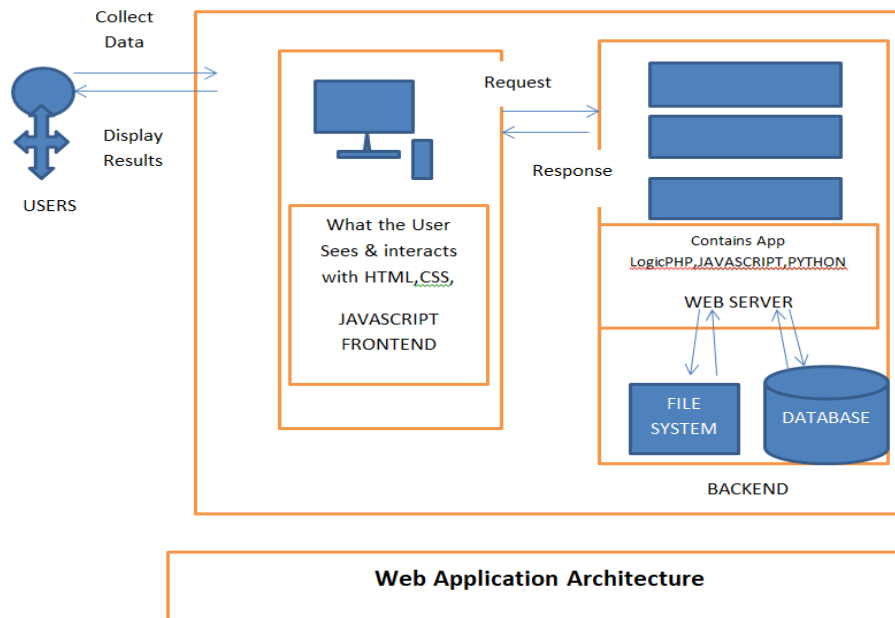


Fig 5.2.1: Diagram of our project architecture

5.3 Impact on Society

The use of a machine learning model to forecast thalassemia will benefit society. Thalassemia is an inherited hemoglobin condition that can be life-threatening but can also be prevented. Assessing the local social economic environment and the amount of public consciousness about Thalassemia is critical for developing important protective efforts. Thalassemia predictions will raise social awareness. In this approach, we can adopt Thalassemia prevention measures. We believe that our prediction model will be extremely beneficial to society.

5.4 Impact on Environment

We think Our trained model is certainly not detrimental to the environment. No chemicals, combustibles, and organic acids are needed to operate this model. Therefore, this model will not have any adverse effects on the environment and biodiversity the use of this model will keep people aware about health condition.

5.5 Ethical Aspects

This Thalassemia prediction model is not anti-moral and does not violate human rights in such a way. The model does not collect any personal information, name, identity etc. so privacy problems will not be occurred. This model plays an important role in making a person aware. The prediction model was created with respect to all types of rules and with respect to privacy and confidentiality issues. So using machine-learning approaches, the prediction model can be managed without any complexity.

5.6 Sustainability Plan

Community, financial and organizational are three parts of Sustainability Plan. The Sustainability Plan gives us a authentic idea of any project to run and future plans for the project. Our model mission is to find the persons who have Thalassemia or not. This model has to be targeted to make it easy for people to adjust and it is important to keep in mind that people do not suffer from inferiority to use this model. Doctors, Diagnostic centres can use this model to speed up their work.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of the Research

Data collection, data preparation, evaluation are all components of the project that we completed. We used the Kaggle dataset to get information. While collecting data, we focused on both patients and non-patients. To boost the quantity of data points, we used the SMOTE approach. We processed and worked on data processing and implementation using Anaconda Navigator and Jupyter Notebook after the data collecting was completed. We run ten machine-learning algorithms after preprocessing, including kNN, LR, Gaussian nave Bayes, RF, ADA boosting, DT, multilayer perceptron, Gradient boosting classifier, and Support vector machine, and evaluate their accuracy, sensitivity, precision, and other metrics. The AdaBoost algorithm, with a 100 percent accuracy, outperforms the others. As a result, we can satisfy the challenge of our model, which is to predict Thalassemia, by utilizing a AdaBoost approach. And Lastly we have implemented a web based project for detecting Thalassemia positive or negative. And also attached a blood bank link through which patients in case of need can save life using that website and.

6.2 Limitations and Conclusions

The goal of our research is to employ machine learning techniques to forecast Thalassemia. Our study and model have several limitations and inconsistencies. The dataset we utilized is too little; it might be a much larger and more vibrant dataset. We were unable to acquire a larger dataset due to several restrictions. For data processing, a variety of advanced approaches could be used, and the model might be presented in a more successful way by using a variety of algorithms. It is possible to determine the symptoms to forecast Thalassemia in a patient using our model. We hope that the general public will adopt our recommended model. After the successful completion of this project one will sense the importance of the project and this project will prominently spread and raise awareness among the among the mankind. It is very crucial to always observe in

order to avoid the risk of Thalassemia. We believe that our model will help the people to predict Thalassemia and to tackle it in a positive manner.

6.3 Implication for Further Study


We simply cannot imagine taking a single step without current science and technology. In the following days, we will upload our program to a platform such as the Android App Store or the Apple App Store, in order to continue the use of information technology and the internet in our country. In the near future, our ongoing efforts in this area will improve the model's accuracy by employing a larger dataset. Furthermore, the model's software can be made accessible to the public by constructing user-friendly GUIs. The current model can be made more effective in the future by implementing new algorithms, introducing additional parameters, and adding some more features. With the assistance of the Thalassemia Department in Bangladesh, our project will be able to achieve greater success and progress. Also we will work on our project further and will try to link a blood donation project with this present project to make it more reliable and user friendly.

REFERENCES

- [1]. M.S. Borah, B. P. Bhuyan, M. S. Pathak, and P. K. Bhattacharya; “Machine Learning in Predicting Hemoglobin Variants,” International Journal of Machine Learning and Computing” vol. 8, no. 2, pp. 140-143, 2018.
- [2]. N. C. Kolambage, H. W. Goonasekara, Dr. R Hewapathirana “Design, Development and Implementation of a Machine Learning-based Predictive Modelling Tool to Accurately Predict Thalassemia Carrier state using Full Blood Count Indices and Haemoglobin Variants” (2020).
- [3]. W. Wongseree, N. Chaiyaratana, K.Vichittumaros, P. Winichagoon, S. Fucharoen. “Thalassaemia classification by neural networks and genetic programming”, 177(3), 771–786. doi:10.1016/j.ins.2006.07.009, (2007).
- [4]. B. Çil, H. Ayyıldız, T. Tuncer, Discrimination of β -Thalassemia and Iron Deficiency Anemia through Extreme Learning Machine and Regularized Extreme Learning Machine Based Decision Support System, Medical Hypotheses (2020), doi: <https://doi.org/10.1016/j.mehy.2020.109611>.
- [5]. Laengsri, V., Shoombuatong, W., Adirojananon, W. et al. “ThalPred: a web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia”, BMC Med Inform Decis Mak 19, 212 (2019). <https://doi.org/10.1186/s12911-019-0929-2>
- [6]. Roberta Risoluti, Stefano Materazzi, Francesco Sorrentino, Carlotta Bozzi, PatriziaCaprari, “Update on thalassemia diagnosis: New insights and methods”, Talanta, Volume 183, 2018, Pages 216-222, ISSN 0039-9140.
- [7]. Fu, Y.-K.; Liu, H.-M.; Lee, L.-H.; Chen, Y.-J.; Chien, S.-H.; Lin, J.-S.; Chen, W.-C.; Cheng, M.-H.; Lin, P.-H.; Lai, J.-Y.; Chen, C.-M.; Liu, C.-Y. “The TVGH-NYCU Thal-Classifer: Development of a Machine-Learning Classifier for Differentiating Thalassemia and Non-Thalassemia Patients”. *Diagnostics* **2021**, *11*, 1725.
- [8]. Aszhari, F. R., et al. "Classification of thalassemia data using random forest algorithm." *Journal of Physics: Conference Series*. Vol. 1490. No. 1. IOP Publishing, 2020.
- [9]. N C Egejuru, S O Olusanya, A O Asinobi, O J Adeyemi, Victor O Adebayo, P A Idowu. “Using Data Mining Algorithms for Thalassemia Risk Prediction. International Journal of Biomedical Science and Engineering”. Vol. 7, No. 2, 2019, pp. 33-44.
- [10]. Alaa S. AlAgha, HossamFaris, Bassam H. Hammo, Ala’ M. Al-Zoubi, “Identifying β -thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine”, Artificial Intelligence in Medicine, Volume 88, 2018, Pages 70-83, ISSN 0933-3657,
- [11]. Farhadi, ShirinDohkt, Mohammad Mehdi Sepehri, and AliAkbarPourfathollah; “The Prediction of Complications of Blood Transfusion in Thalassemia Patients Using Deep Learning Method”; International Journal of Hospital Research 7.4 (2018): 116-130.
- [12]. P. Paokanta, N. Harnpornchai, S. Srichairatanakool, and Ceccarelli, “The Knowledge Discovery of [beta]-Thalassemia Using Principal Components Analysis: PCA and Machine Learning Techniques”; International Journal of e-Education, e-Business, e-Management and e-Learning 1.2 (2011): 169.

- [13]. P. Paokanta, M. Ceccarelli and S. Srichairatanakool, "The efficiency of data types for classification performance of Machine Learning Techniques for screening β -Thalassemia", 2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2010), 2010, pp. 1-4, doi: 10.1109/ISABEL.2010.5702769.
- [14]. Sacco, Massimiliano and Sciandra, Mariangela and Maggio, Aurelio, "Random Forest Analysis: A New Approach for Classification of Beta Thalassemia" (March 19, 2020). d/SEAS.
- [15]. Jahangiri, Mina; Khodadi, Elahe; Rahim, Fakher; Saki, Najmaldin; Saki Malehi, Amal (2017). "Decision-tree-based methods for differential diagnosis of β -thalassemia trait from iron deficiency anemia". Expert Systems, (), e12201–doi:10.1111/exsy.12201
- [16]. M. S. Borah, P. K. Bhattacharya, M. S. Pathak, and D. Kalita, "A hospital based study of Hb variant and beta thalassaemia mutational pattern characterization among the people of Northeast region of India," Annals of Pathology and Laboratory Medicine, vol. 3, no. 3, pp. 134-140, 2016.
- [17]. S. L. Thein and J. Rochette, "Disorders of hemoglobin structure and synthesis," Principles of Molecular Medicine, NJ: Humana Press, 1998, pp. 179-190.
- [18]. M. S. Pathak, M. S. Bora, and D. Kalita, "Disorders of haemoglobin variants in paediatric patients attending in a tertiary care hospital of North East India," International Journal of Biological and Medical Research, vol. 5, no. 1, pp. 3841-3846, 2014.
- [19]. M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9, no. 1, pp.1-16, 2017.
- [20]. N. Esfandiari, M. R. Babavalian, A. E. Moghadam, and V. K. Tabar, "Knowledge discovery medicine: Current issue and future trend," Expert Systems with Applications, vol. 41, no. 9, pp. 4434-4463, 2014.
- [21]. S. R. Amendolia, G. Cossu, M. L. Ganadu, G. Bruno, G. L. Masala, and G. M. Mura, "A comparative study of k-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening," Chemometrics and Intelligent Laboratory Systems, vol. 69, no. 1, pp. 13-20, 2003.
- [22]. W. Wongseeree, N. Chaiyaratana, K. Vichittumaros, P. Winichagoon, and S. Fucharoen, "Thalassaemia classification by neural networks and genetic programming," Information Sciences, vol. 177, no. 3, pp. 771-786, 2007.
- [23]. E. A. El-Sebakhy and M. A. Elshafei, "Thalassemia Screening Using Unconstrained Functional Networks Classifier," in Proc. IEEE International Conference on Signal Processing and Communications, Dubai, United Arab Emirates, 2007, pp. 1027-1030.

PLAGIARISM REPORT

Plagiarism Checked by Abdus Sattar, Assistant Professor, Department of CSE	 28-12-2021
--	--

THALASSEMIA PREDICTION USING MACHINE

ORIGINALITY REPORT

19%

SIMILARITY INDEX

12%

INTERNET SOURCES

10%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1

dspace.daffodilvarsity.edu.bd:8080

Internet Source

6%

2

Submitted to Daffodil International University

Student Paper

3%

3

Betül Çil, Hakan Ayyıldız, Taner Tuncer.

"Discrimination of β -thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system", Medical Hypotheses, 2020

Publication

1%

4

Md. Ariful Islam Arif, Saiful Islam Sany, Farah Sharmin, Md. Sadekur Rahman, Md. Tarek Habib. "Prediction of addiction to drugs and alcohol using machine learning: A case study on Bangladeshi population", International Journal of Electrical and Computer Engineering (IJECE), 2021

Publication

1%

5

Waranyu Wongseeree, Nachol Chaiyaratana, Kanjana Vichittumaros, Pranee Winichagoon,

1%