

Thalassemia Prediction using Machine Learning Model

BY

Md Golam Rabbani

ID: 181-15-1808

Sharmila Zaman

ID: 181-15-1796

Reaz Uddin Hemel

ID: 181-15-1739

The report is presented in the Partial Fulfillment of the demand for the completion of B.Sc. in Computer Science and Engineering

Supervised By

Al Amin Biswas

Lecturer (Senior Scale)

Department of CSE

Daffodil International University

Co-Supervised By

Saima Afrin

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Project titled “**Thalassemia Prediction using Machine Learning Model**”, submitted by Md Golam Rabbani, Sharmila Zaman, and Reaz Uddin Hamel to the Department of CSE. The report was accepted by Daffodil International University as up to the mark for the partial fulfillment of the necessity for the completion of B.Sc. in CSE and received as to its content. The presentation will be held on 13.01.2022.

BOARD OF EXAMINERS



Mohammad Monirul Islam
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Mahfujur Rahman [MMR]
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Swakkhar Shatabda
Associate Professor,
Department of Computer Science and Engineering (CSE)
United International University (UIU), Dhaka, Bangladesh

External Examiner

DECLARATION

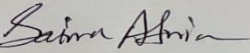
We herewith state that the research has been done by us under the direct supervision of **Al Amin Biswas**, Lecturer (Senior Scale), Department of CSE, Daffodil International University. We proclaim that this report is unique and written by us and not submitted anywhere else.

Supervised by:



Al Amin Biswas
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised by:



Saima Afrin
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Md Golam Rabbani
ID: 181-15-1808
Department of CSE
Daffodil International University



Sharmila Zaman
ID: 181-15-1796
Department of CSE
Daffodil International University



Reaz Uddin Hemel
ID: 181-15-1739
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

Firstly, we are delighted to convey our gratitude to the almighty for his divine blessings in making us possible to accomplish the research-based project of the final year. We are appreciative to **Al Amin Biswas**, Lecturer (Senior Scale) of the Department of CSE Daffodil International University for his contribution and supervision in this research-based project by his knowledge and interest in the field of “Machine Learning”. His constant support, patience, guidance, and advice helped us to implement this project in real. We would like to convey our gratitude to **Al Amin Biswas**, Lecturer (Senior Scale) and **Saima Afrin**, Lecturer, Department of CSE and other faculty members of the CSE department of DIU for their guidance to complete our research-based project. We are grateful to our entire team in DIU, who attended this discussion while implementing the work. Finally, we must admit the support and patients of our parents with due respect.

ABSTRACT

Thalassemia is a genetic blood disease inherited from parents. It is the most common and concerning genetic disorder globally. Minor to major anemia and transfusion dependence is the main symptom of this disease. In South Asian countries like Bangladesh, every year there are many children born with thalassemia traits. Among various types of thalassemia, beta-thalassemia is the most severe one that causes weakness, serious anemia, shortness of breath, even failing organs like the kidney, heart. This study aims to classify thalassemia depending on the values of various hemoglobin (Hb) indices like Hb A, Hb B, Hb E, and Hb F collected from the data of a thalassemia center of Bangladesh. This work is to depict the epidemiological aspects of thalassemia from the data of the common people of all stages of Bangladesh. We applied various machine learning classifiers such as Logistic Regression (LR), Decision Tree, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN), etc. to classify thalassemia. For evaluating the performance of the classifiers, we calculated accuracy, precision, recall and f1-score. We also plotted the ROC curve. From the ROC curve, it is observed that AUC (Area Under the Curve) has a big area. After conducting the study, we got the final result that concludes that among all the algorithms, the Random Forest and K-Nearest Neighbors (KNN) have shown the best accuracy which is 99.14%. The both precision and recall for the Random Forest is 99.00% and for KNN is 99.00% and 100% respectively.

Keywords: Thalassemia, Confusion matrix, Classifier, ML Models, Random Forest.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval	i
Declaration	ii
Acknowledgment	iii
Abstract	iv
CHAPTERS	
CHAPTER 1: INTRODUCTION	01-02
1.1 Introduction	01
1.2 Motivation	02
1.3 Challenge	02
1.4 Objective	02
1.5 Report Organization	02
CHAPTER 2: LITERATURE REVIEW	03-04
2.1 Literature Review	03-04
CHAPTER 3: METHODOLOGY	05-12
3.1 Data Collection	05
3.2 Data Preprocessing	06-07
3.3 Model Description	08-12
3.4 Model Training & Testing	12
CHAPTER 4: RESULT ANALYSIS	14-19
4.1 Confusion Matrix	13-14
4.2 Result Analysis	15-16
4.3 ROC Curve	16-17
4.4 Comparative Analysis	18-19

CHAPTER 5: CONCLUTION AND FUTURE WORK	20-22
5.1 Conclusion	20
5.2 Future Work	20
5.3 Acknowledgment	20
References	21-22

List of Figures		
Figure No	Figure title	Page
Figure 01	Bar Diagram of our Dataset	06
Figure 02	Implementation Procedure to Predict the Thalassemia	07
Figure 03	Logistic Regression	08
Figure 04	K-Nearest Neighbors	09
Figure 05	Decision Tree Classifier	10
Figure 06	Support Vector Machine	11
Figure 07	The General Illustration of Random Forest	12
Figure 08	ROC Curve for selected models	17

List of Tables		
Table No	Title of Table	Page No
Table 01	Variables for Thalassemia Prediction	06
Table 02	Confusion Matrix Based on the Test set	13-14
Table 03	Experimental Result of Applied Machine Learning Models	16
Table 04	Comparison Our work Among the Other's works.	18-19

CHAPTER 1

INTRODUCTION

1.1 Introduction

Thalassemia is one of the most outgoing global health concerns. It is a very common genetic disorder, especially in southeast countries of Asia, Africa, and the Middle East. It is an inherited hemoglobin disorder caused by abnormal structure or production of hemoglobin in red blood cells. Inherited means it passes from parents to children through genes. Hemoglobin is made of two types of Protein, such as Alpha, and Beta that transports oxygen to all body cells and tissues. This disorder results in the red blood cells not being enough to produce hemoglobin or being destroyed, which leads to anemia. Thalassemia is the most common disease of hemoglobin. Every year over 330,000 babies are born with this disease. It has been indicated as a growing health concern in most countries by WHO. 280 million people were affected with thalassemia disorder with about 439,000 having severe symptoms as of 2015. In the same year, it caused 16,800 deaths. In this research, a dataset containing four types of hemoglobin (Hb) indices Hb A, Hb B, Hb E, and Hb F, of 1735 patients' data is used to train models of machine learning. This study is performed on various models like the random forest, Gaussian Naïve Bayes, SVM, K-NN. We have tried to predict thalassemia disease using machine learning approaches. The whole report is arranged as follows: The literature review of thalassemia-related papers is described in Chapter 2. The methodology is explained in Chapter 3 with data collection, preprocessing of data, model description, and training and testing. Next, the final analysis of Results is described in Chapter 4. Finally, the last Chapter 5, concludes the summarization of this research and our future work.

1.2 Motivation

The purpose of this study is to classify thalassemia from various Hemoglobin (Hb) indices like Hb A, Hb B, Hb E, and Hb F. This work is to depict the epidemiological aspects of thalassemia from the data of the common people of all stages of Bangladesh.

1.3 Scope & Challenges

In Bangladesh, more than 14000 children are born thalassemia every year and 10 percent of the total population are carrying the disease. Most of them don't even know about it. That's why an early and easy prediction can be an advantage. By this work, a new research path has been created that would be helpful to find the prevention of Thalassemia. Thoroughly collecting and working with such a large and imbalanced dataset, was a little challenging initially. It was being challenge for us to gather real-time pathological data because of privacy issues

1.4 Objectives

This study has been done to predict thalassemia by using multiple machine learning techniques. Our focus is to find the best-performing model for our dataset. We collected real-life patient data from a specialized thalassemia hospital and conducted this research.

1.5 Report Organization

Chapter 1: In this part we investigated the Introduction of Thalassemia, Motivation, Scopes and Challenges, Objectives Report Organization.

Chapter 2: Literature review.

Chapter 3: In this chapter, we have Data collection, Data preprocessing, balancing the imbalanced dataset, Implementation procedure to predict the Thalassemia and Model description.

Chapter 4: Confusion matrix, Classification Report, Result Analysis, and ROC Curve.

Chapter 5: Conclusion, Future work, Acknowledgment, and References.

CHAPTER 2

LITERATURE REVIEW

2.1 Literature Review

Aszhari et al. [1] collected data on thalassemia and performed a study of the patients. They proposed a random forest method that can classify the disease thalassemia precisely.

According to their research, the random forest algorithm achieved 100% accuracy, precision, and recall. P. Prokanta [2] conducted research where they tried to propose a machine learning technique of (PCA) for screening a genotype of Beta-Thalassemia patients. The goal of the research was to minimize the dimension of any data before classifying them. Among different algorithms, Multi-Layer Perceptron (MLP) outperformed all other techniques and achieved an accuracy of 86.61%. In this paper, P. Paskenta et al. [3] performed a study comparing the performance of different classification techniques of different kinds of data types to find the most suitable data type of a technique. By using the data of β -Thalassemia patients, they brought the highest accuracy of 85.83% and 84.25% respectively for Multinomial Logistic Regression and Bayesian Networks (BNs) in Nominal data types. Apart from that, they also found an accuracy of 88.98% for KNN, 87.40% for Multi-Layer Perceptron, 84.25% for Naive Bayes with Interval scale. S. Thakur and S.N. Raw [4] used the fuzzy inference system to predict the severity of thalassemia disease. They have designed a mathematical model and displayed using fuzzy rules that showed the different stages of thalassemia of a patient. In this paper, S. Hossain et al. [5] studied over the 2009-2014 time period with 1178 cases of different specialized thalassemia centers to portray the mutation spectrum, epidemiology, clinical course, and treatment outcomes of Bangladesh. They have discussed the preventative strategies of thalassemia and the overall treatment strategies prevailing in Bangladesh. The report - done by R. Das et al. [6], where two separate mechanisms have been used for scoring, such as BTT detection of Hb Reassembles (HbE) trait and BTT. The research resembles the data collected from a medical research center in India. They have worked on two individual data sets and got the accuracy of 79.25% and 91.74% for BTT, 58.62%, and 78.03% for the HbE and BTT, respectively. E. R. Susanto et al. [7] have studied a new model based on fuzzy rules to predict and classify thalassemia for children based on CBC data like the

values of Hb, MCV, and MCH along with its four output models called major, minor, intermediate, and not thalassemia. In this report, a new machine learning classifier has been generated by Yi-Kai Fu et al. [8] discriminating thalassemia and non-thalassemia microcytic anemia. After studying about 350 Taiwanese thalassemia patients, they generated the classifier using a Support Vector Machine (SVM) which showed a performance of average AUC of 0.76. S Sadiq et al. [9] conducted a study with the data of Punjab Thalassemia Prevention Project Lab to detect Beta-Thalassemia carriers from their blood test. They have proposed a new model combining three machine learning models such as Support Vector Machine (SVM), and Random Forest, and Gradient Boosting Machine that has shown an accuracy of 93.00%.

CHAPTER 3

METHODOLOGY

This part has been designed to present the methodology of our work. It has been separated into four parts. 3.1. Data Collection, 3.2. Data Preprocessing, 3.3. Models Description, 3.4. Model training and testing. Dataset has been deeply described in the 3.1 and 3.2 subsections respectively and 3.3 and 3.4 subsections include descriptions of different kinds of machine learning models for prediction and analysis.

3.1 Data Collection

To complete this study, 1733 real-time data points were collected from the Filariasis General Hospital and Thalassemia Institute, Zinzira, Savar, Dhaka, Bangladesh, from April 16, 2018 to May 13, 2021. This dataset also includes data from university students, garment workers, and the general public.

3.2 Data Preprocessing

The independent variables of this dataset are Hb A, Hb F, Hb E, and Hb A2. The target attribute is a type of Thalassemia and Carrier. This dataset includes 1448 Normal and 285 Positive records. A brief description has been given about the attributes of the dataset in Table01 and the Graphical view is shown in Figure 01. To order to preprocess our dataset we had to Handle the Null value, Missing data handle and remove the noisy values etc.

3.2.1 Balance the Imbalanced Dataset

There are six classes in our dataset with few imbalanced data distributions. That is why we had to balance the classes by using SMOTE technique. SMOTE is a machine learning technique that helps to balance the imbalanced dataset by creating sample cases from the existing data to make the dataset reliable.

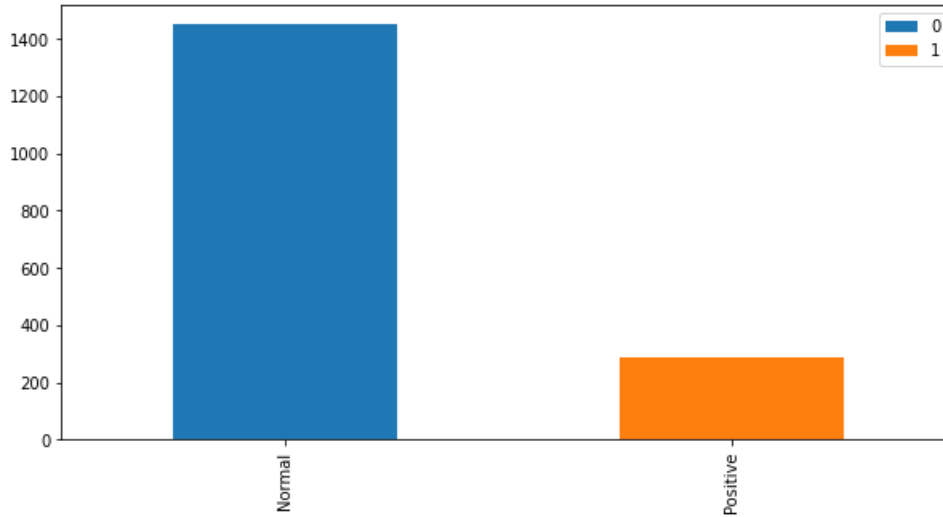


Fig. 01: Bar Diagram of our Dataset

Table 01: Variables for Thalassemia Prediction.

Variable Name	Variable Type	Variable Description	Possible Values
Hb A	Predictor	Part of Hemoglobin	0 - 100
Hb F	Predictor	Part of Hemoglobin	0 - 100
Hb E	Predictor	Part of Hemoglobin	0 - 100
Hb A2	Predictor	Part of Hemoglobin	0 - 100
Class	Target	It is a target variable that predicts thalassemia patient, carrier, and normal stage.	Normal (0), Positive (1)

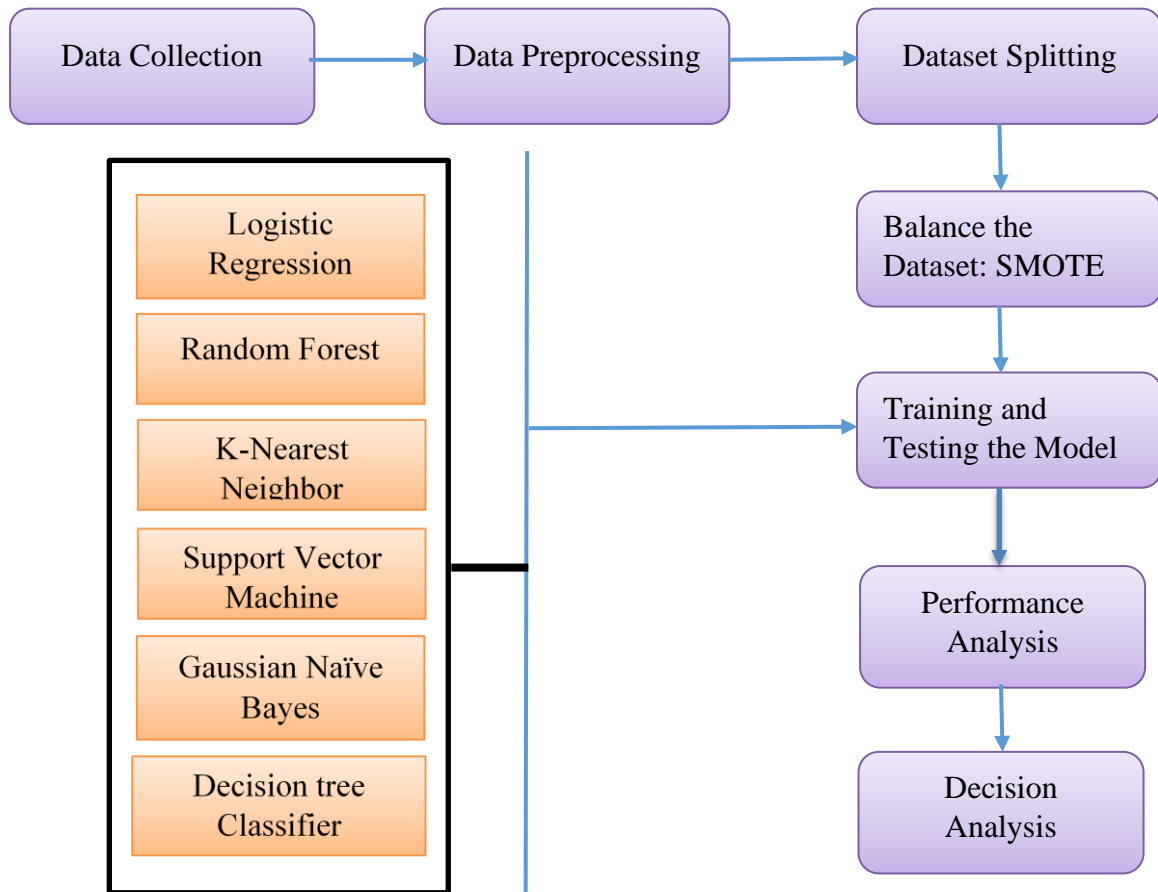


Fig.02: Implementation Procedure to Predict the Thalassemia

3.3 Model Description

3.3.1 Logistic Regression: A technique that can be used for traditional statistics and predicting the dependent variable by using a group of independent variables. Logistic regression predicts the outcome of traditional statistics of a dependent variable. Hence the output has to be a discrete value. It may be '0' or '1', or true or false or yes or no, et cetera. But it does not give the exact value, it gives the possibility value which lies between '0' and '1'. In Logistic regression, there is a function called sigmoid that converts the independent variable into an expression. It has the probability with ranges of '0' and '1' for the dependent variable. In Logistic regression, we fit an S-shaped logistic function to determine the maximum value between '0' and '1'. By '0' we get no possibility of occurrence and by '1' we get the possibility of occurrence.

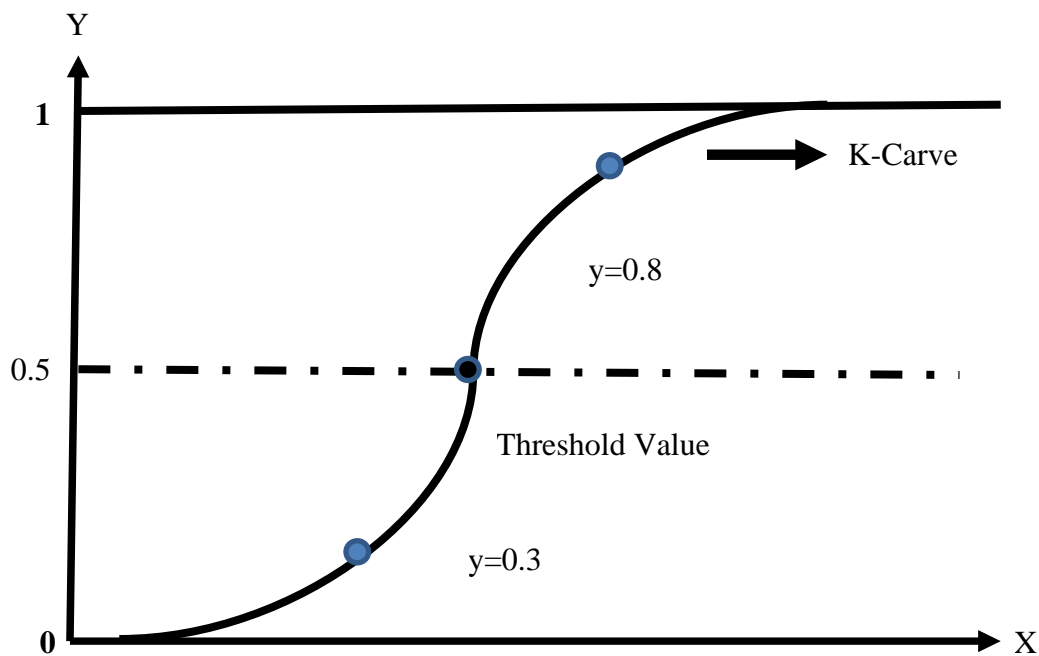


Fig.03: Logistic Regression

3.3.2 KNN Algorithm: KNN stands for K-Nearest Neighbor Algorithm, which is a classifier and is a supervised learning algorithm. The new and available data are used to function KNN, which inserts the new data into a category that resembles the other available categories. Although the K-NN technique can be used for regression, it is most commonly used for classification issues. The K-Nearest Neighbor algorithm is a non-parametric method. It can't be used with historical data. Assume there are two categories, M and N, and a data item A1 that can fall into any of them. We can use the KNN to solve this problem. We can easily classify any dataset by employing the KNN method.

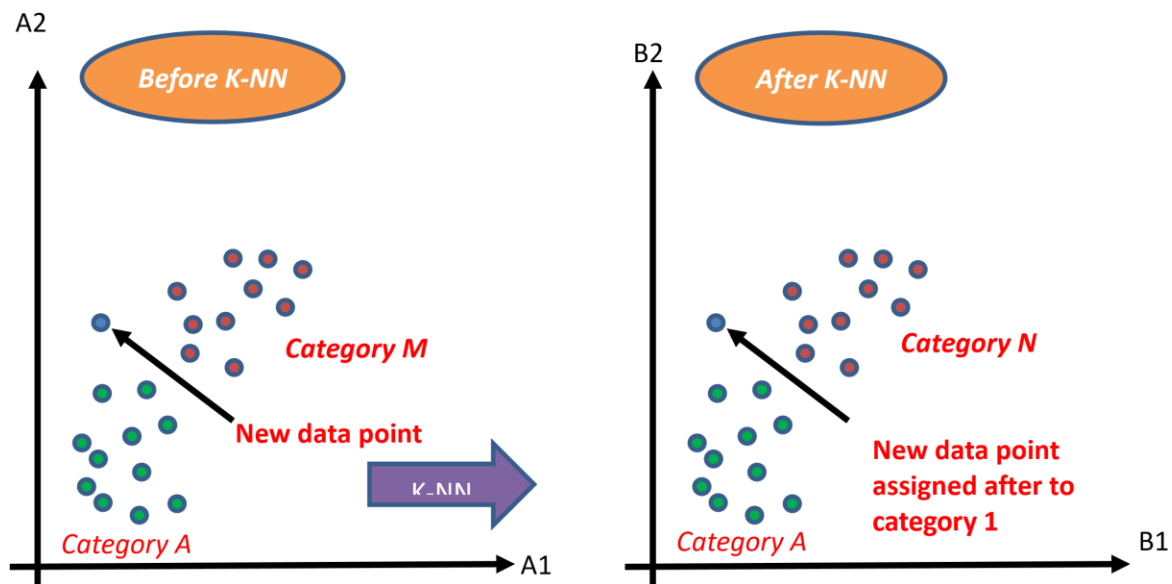


Fig.04: K-Nearest Neighbors

3.3.3 Decision Tree Classifier: For classification and regression, a decision tree is utilized. Two nodes, such as 'Decision Node' and 'Leaf Node,' make up a Decision Tree. The selected nodes are used to make any kind of decision and can have several branches, with Leaf nodes serving as the Decisions' output. The Decision Tree is based on ranked data, with 'yes' indicating a high likelihood of occurrence and 'no' indicating a low likelihood of occurrence. It looks like a tree, which is why it's called a decision tree, and it starts at the root node and branches out to form an algorithm that looks like a tree.

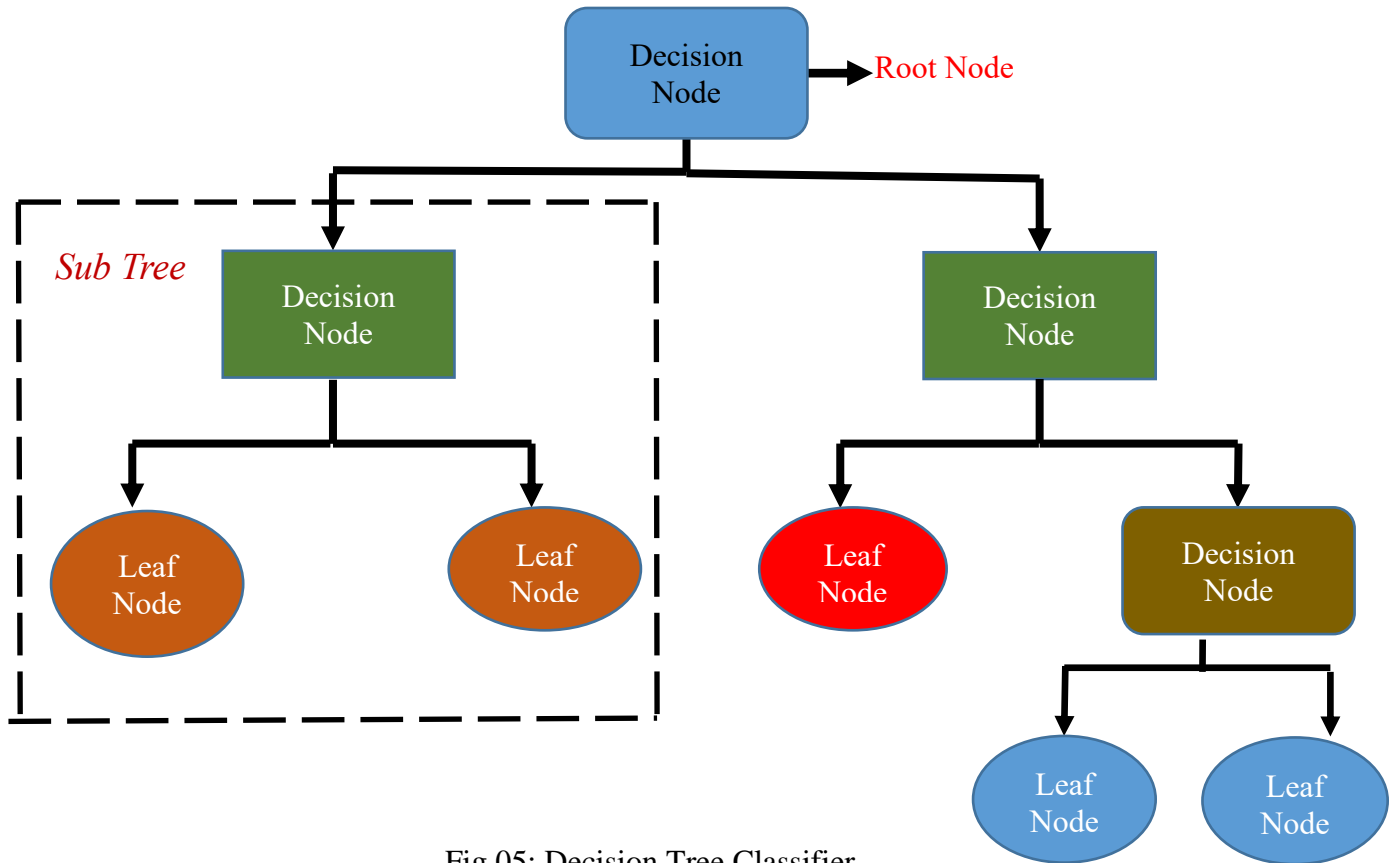


Fig.05: Decision Tree Classifier

3.3.4 Support Vector Machine (SVM): Supervised Learning (SVM) is a of supervised learning. It's also known as SVM, Due to Supervised Learning, it is dependent on the level data approach. It's utilized for regression analysis and categorization. It's also utilized to solve classification issues. New data is always assigned to classes, and the best decision boundary or hyper plane is predicted. SVM is utilized for this reason. The Support Vector Machine selects extreme points to aid in the creation of hyper planes.

Linear SVM: Linear Classifier is used for data that are linearly separable, which means that two datasets can be divided into classes using a straight line that serves as the hyper plane or decision boundary. These datasets are referred to as linearly separable data, and the classifier is used for Linear Classifier.

Non-linear SVM: This one is applied for Non-Linear separable data, which means datasets cannot be separated and cannot be classified by using a single straight line.

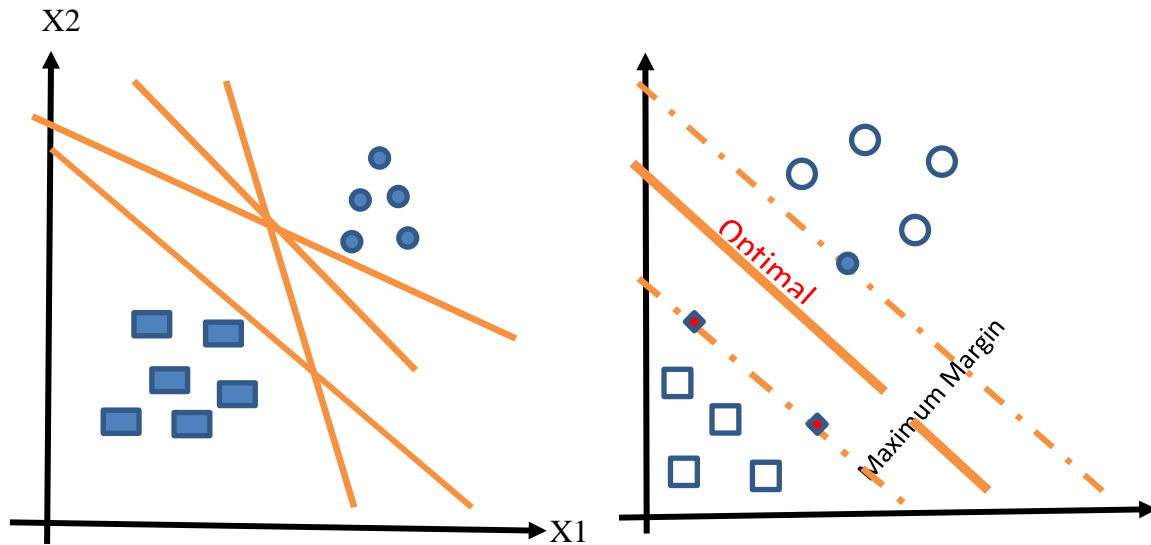


Fig.06: Support Vector Machine.

3.3.5 Gaussian Naive Bayes

Gaussian Naive Bayes cannot be used in discrete data. But we can use Gaussian Naive Bayes in continuous feature variables or data. It is an alternative to NB which adopts continual data and Gaussian normal distribution. It is a group of supervised learning algorithms which is used for classifying. It is based on the Bayes theorem.

$$\text{PDF} = P(x_i | y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \dots \dots \dots (1)$$

By this PDF formula, we can search out mean, variance, and standard deviation to implement Gaussian Naive Bayes.

3.3.6 Random Forest

It is one of the most popular algorithms that belong to supervised learning. Random forest is built up from decision trees. It is used for classifications and regression problems. In a random forest, a decision tree is used with many subsets of the data set that helps to get the accuracy. We have to consider two datasets. Original dataset and Bootstrap dataset. Then the original data set will be assigned in the bootstrap dataset by selecting randomly. To make a random forest, we need to select nodes from Bootstrap Dataset by considering a subset of variables at each step.

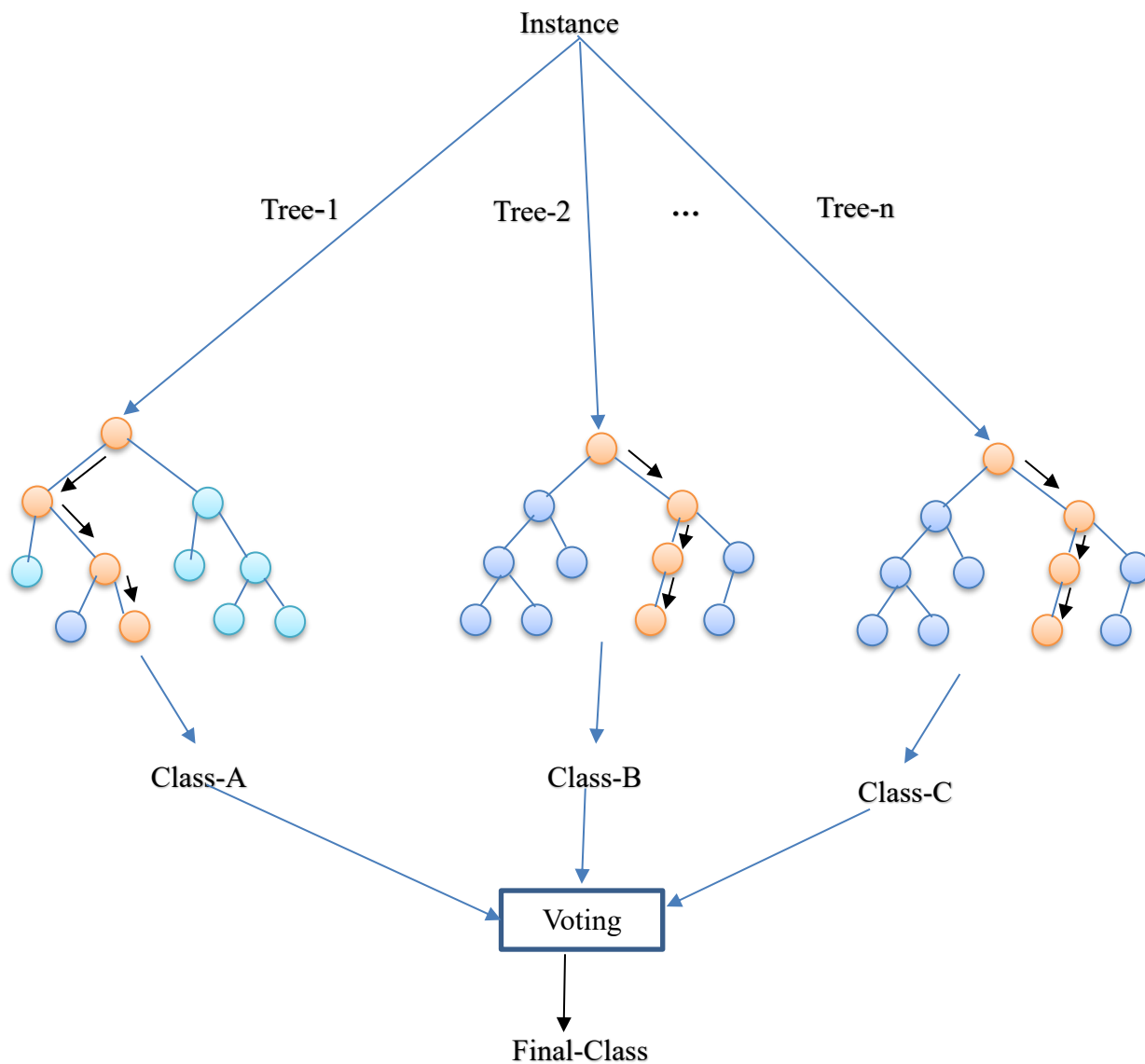


Fig.07: Working Flow of Random Forest

3.4 Model Training & Testing

To train and test the six distinct Models, 80 percent (80.00%) of the data was used to train, which means 1388 records from the dataset were used to train, and 20 percent (20.00%) of the data was used to test, which means 347 records from the dataset were used to test the accuracy.

CHAPTER 4

RESULT ANALYSIS

4.1 Confusion Matrix

A confusion matrix is a matrix by which a visual representation of the Actual vs Predicted values is shown by $n*n$ matrix. It is a major thing to evaluate a machine learning model.

True Positive (TP): The number of instances that were positive and correctly classified as positive.

False Positive (FP): The number of instances that were negative and incorrectly classified as positive.

True Negative (TN): The number of instances that were negative and correctly classified as negative.

False Negative (FN): The number of instances that were positive and incorrectly classified as negative.

Table 02: Confusion Matrix Based on the Test set

Random Forest		Predicted Class	
		Normal	Positive
Actual Class	Normal	287	3
	Positive	2	287
Decision Tree		Predicted Class	
		Normal	Positive
Actual Class	Normal	287	3
	Positive	5	284

K-Nearest Neighbors (KNN)		Predicted Class	
		Normal	Positive
Actual Class	Normal	286	4
	Positive	1	288
Logistic Regression		Predicted Class	
		Normal	Positive
Actual Class	Normal	287	3
	Positive	18	271
Support Vector Machine (SVM)		Predicted Class	
		Normal	Positive
Actual Class	Normal	287	3
	Positive	20	269
Naive Bayes		Predicted Class	
		Normal	Positive
Actual Class	Normal	289	1
	Positive	37	252

4.2 Result Analysis

Accuracy: Accuracy is the fraction of the number of examples that are correctly classified by the classifier to the total number of instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots (2)$$

Precision: Precision is the degree by which a process will repeat the same value.

$$\text{Precision} = \frac{TP}{TP+FP} \dots\dots\dots (3)$$

True Positive Rate (TPR)/Recall: It is defined as the function of the positive examples predicted correctly by the classifier.

$$\text{Recall} = \frac{TP}{TP+FN} \dots\dots\dots (4)$$

False Positive Rate (FPR): It is defined as the function of negative examples classified as a positive class by the classifier.

$$\text{FPR} = \frac{FP}{TN+FP} \dots\dots\dots (5)$$

False Negative Rate (FNR): It is defined as the function of positive examples classified as a negative class by the classifier.

$$\text{FNR} = \frac{FN}{FN+TP} \dots\dots\dots (6)$$

True Negative Rate (TNR): It is defined as the function of negative examples classified correctly by the classifier. It is also called specificity.

$$\text{TNR} = \frac{TN}{TN+FP} \dots\dots\dots (7)$$

F1-Score: Recall and Precision are two widely used metrics employed in the analysis.

$$\text{F1-Score} = \frac{2TP}{2TP+FP+FN} \dots\dots\dots (8)$$

Table 03: Experimental Result of Applied six ML Models

Model	Accuracy	Precision	Recall	F1-score
Random Forest	99.14%	99.00%	99.00%	99.00%
KNN	99.14%	99.00%	100.00%	99.00%
SVM	96.03%	99.00%	93.00%	96.00%
Nave Bayes	93.44%	100.00%	87.00%	93.00%
Decision Tree	98.62%	99.00%	98.00%	99.00%
Logistic Regression	96.37%	99.00%	94.00%	96.00%

In this above table we can see that Random Forest and K-Nearest Neighbors have the best Accuracy and Precision which is 99.14% and 99.00% respectively. Here we consider only positive class because we have been considered the positive class as True Positive.

4.3 ROC Curve

A ROC curve is an evaluation matrix or a binary classifier system that defines the performance of a classifier. It shows the distinguishment between the classes. ROC curve is calculated by the rate True positive and False Positive against each other. AUC or the area under the curve is used to measure the ability of the classifier as a summary of the ROC curve.

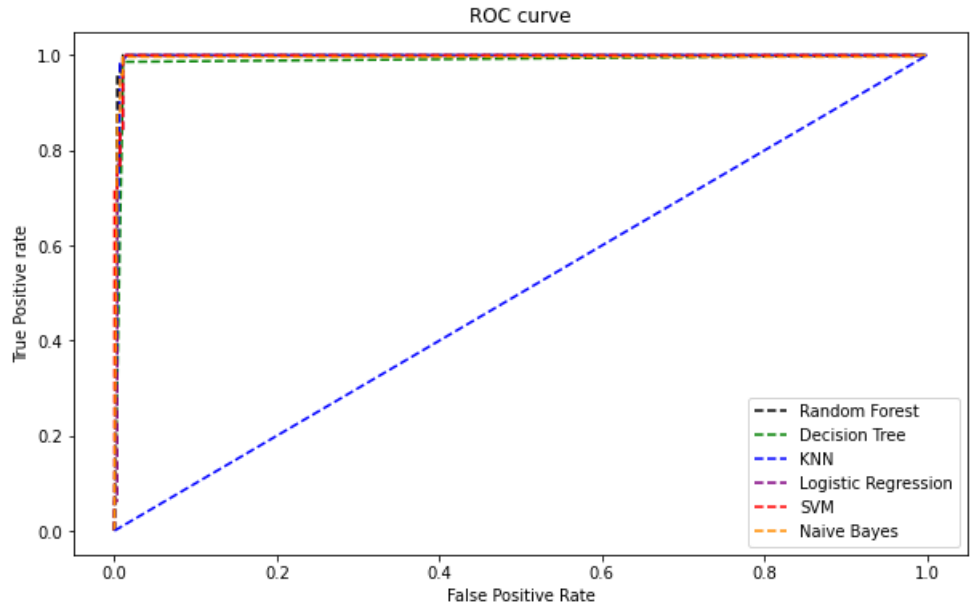


Fig08: ROC Curve for the selected models

4.4 Comparative Analysis

In this section we compared our work among the other's works and we made a table for this. The table is given below.

Table 04: Comparison Among Our work and the Others work

Comparison Table					
Serial No.	Author	ML Model Used	Prediction	Accuracy	
01	Our work	RF, DT, KNN, LR, SVM, NB	Prediction of Thalassemia	RF (99.14%), DT (98.62%), KNN (99.14%), LR (96.37%), SVM (96.03%), NB (93.44%)	
02	F R Aszhari	RF	Classification of Thalassemia	RF (100.00%)	
03	Patcharaporn Paokanta	MLP, KNN, NB, BNs, MLR	Classification of Thalassemia & feature selection by PCA	MLP (86.00%), KNN (85.00%), NB (85.00%), BNs (92.00%), MLR (82.00%)	
04	Patcharapom Paokanta	KNN, MLP, NB, BNs, MLR	Compare Data Types for Screening β -Thalassemia	NOMINAL SCALE (KNN (88.98%), MLP (87.40%), NB (84.25%), BNs (83.46%), MLR (81.89%))	INTERVAL SCALE (KNN (88.98%), MLP (85.83%), NB (83.46%), BNs (85.83%), MLR (84.25%))
05	S. Thakur	Mamdani Fuzzy Inference System	Thalassemia risk prediction	Satisfactory	
06	M S Hossain		Clinical Lesson of Thalassemia from BD	DT (85.25%), NB (90.74%)	

07	Reena Das	DT, NB, ANN	Identify BTT & differentiate normal vs carrier	For BTT DT (79.25%), NB (91.74%) DT (58.62%), NB (78.03%)
08	E. R. Susanto	Fuzzy Based Model	Identify the type of Thalassemia	Worked well
09	Yi-Kai Fu	SVM	Discriminating thalassemia and non-thalassemia	SVM (76.00%)
10	S Sadiq	SGR-VC (SVM, GBM and RF)	Classification of β -Thalassemia Carriers	SGR-VC (93.00%), (SVM (90.00%), GBM (91.00%) and RF (91.00%))

From the above analysis we came to the point that, Random Forest 99.14%, Decision tree classifier 98.62 %, K-Nearest Neighbors 99.14.00 %, Logistic Regression 96.37.00 %, Support Vector Machine 96.03% and Gaussian Nave Bayes 96.44%. The highest accuracy is 99.14%. Random Forest and K-Nearest Neighbors provided the maximum accuracy which is 99.14%. As a result, Random Forest and K-Nearest Neighbors are the best classifier models for the dataset. In the future, we aim to expand the data properties and use a larger dataset to predict thalassemia.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

We used multiple machine learning classifiers to detect and classify the forms of thalassemia in this paper. We obtained real-time data from a reopened thalassemia hospital in Bangladesh for various age groups and genders. We used 1388 data for training and 347 data for testing, accounting for 80.00% of the total dataset. To categorize the different kinds of thalassemia. For classification, we employed Random Forest, Decision Tree Classifier, K-nearest Neighbors, Logistic Regression, and Support Vector Machine machine learning models. All of the classifiers have performed admirably. However, the Random Forest and K-Nearest Neighbor classifiers performed the best, with a 99.14% accuracy. After evaluating the findings, it can be concluded that the Random Forest and K-Nearest Neighbors models are both very efficient in Thalassemia classification and outperform all other models used. In this case, we looked at different hemoglobin indices in the blood test data. The primary goal of this research is to evaluate the effectiveness of machine learning models for categorization tasks.

5.2 Future Work

The main focus of this research is on analyzing the performance of machine learning models for categorization tasks. In the future, we'll aim to expand the data properties and use a larger dataset to predict thalassemia.

Acknowledgment: We would like to acknowledge and thank Professor Dr. Moazzam Hossain, MBBS, and M.Phils. (Micro-Biology, Bio-Chemistry, and Bio-Technology), Ex-director disease control DGHS chairman of Filaria General Hospital and Thalassemia Institute, Birulia, Savar, Dhaka, Bangladesh, for sharing the dataset for the research.

References

- [1] F R Aszhari, Z Rustam, F Subroto, and A S Semendawai, “*Classification of Thalassemia data using Random Forest algorithm*”, Journal of Physics: Conference Series, Volume 1490, 2019
- [2] P Paokanta, N Harnpornchai, S Srichairatanakool, and M Ceccarelli, “*The knowledge discovery of β -Thalassemia using principal components analysis: PCA and Machine Learning techniques*”, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 1, No. 2, June 2011
- [3] P Paokanta, M Ceccarelli, and S Srichairatanakoo, “*The Efficiency of data types for classification performance of Machine Learning techniques for screening - Thalassemia*”, 978-1-4244-8132-3/10/\$26.00 ©2010 IEEE
- [4] S. Thakur* and S.N. Raw, “*Thalassemia risk prediction model using inference system: An application of Fuzzy Logic*”, Research Journal of Mathematical and Statistical Sciences Vol. 5(7), 1-8, July (2017)
- [5] M S Hossain et al., “*Thalasseмии in South Asia: Clinical Lessons learned from Bangladesh*”, Orphanet Journal of Rare Diseases (2017)
- [6] R Das et al., “*A Decision Support Scheme for Beta-Thalassemia and HbE Carrier Screening*”, Journal of Advanced Research Volume 24, July 2020, Pages 183-190
- [7] E. R. Susanto, A. Syarif, K Muludi, R. R. W. Perdani, A. Wantoro, “*Implementation of Fuzzy-based model for prediction of Thalassemia diseases*”, E. R. Susanto et al 2021 J. Phys.: Conf. Ser. 1751 012034
- [8] Yi-Kai Fuet al., “*The TVGH-NYCU Thal-Classifer: Development of a Machine Learning classifier for differentiating Thalassemia and Non-Thalassemia Patients*”, Diagnostics 2021, 11, 1725. <https://doi.org/10.3390/diagnostics11091725>
- [9] S SADIQ et al., “*Classification of β -Thalassemia carriers from red blood cell Indices using ensemble classifier*”, Digital Object Identifier 10.1109/ACCESS.2021.3066782
- [10] Z Rustam, and R Hidayat, “*Comparison of Fuzzy C-Means, Fuzzy Kernel C-Means, and Fuzzy Kernel Robust C-Means to classify thalassemia data*”, August 2019 International Journal on Advanced Science Engineering and Information Technology 9(4):1205

- [11] Hossain, M.S., Mahbub Hasan, M., Petrou, M. *et al.* The Parental Perspective of thalassemia in Bangladesh: Lack of Knowledge, Regret, and Barriers. *Orphanet J Rare Dis* 16, 315 (2021). <https://doi.org/10.1186/s13023-021-01947-6>
- [12] Centers for Disease Control and Prevention, available at <<<https://www.cdc.gov/ncbddd/thalassemia/facts.html>>> , last accessed on 10-01-2022 at 10.25 PM.