THALASSEMIA PREDICTION USING HYBRID MACHINE LEARNING APPROACHES

NitheshKanna S [1], Mahavishnu G [2], Mithhull B [3]

[1, 2, 3] UG – B. E Electronics and Communication, Bannari Amman Institute of Technology, Sathyamangalam, Tamil Nadu

nitheshkanna.ec20@bitsathy.ac.in, mahavishnu.ec20@bitsathy.ac.in, mithhullbalaji.ec20@bitsathy.ac.in, ramasamis@bitsathy.ac.in

ABSTRACT

This study proposes a workflow for preparing and training a neural network model to classify medical data related to thalassemia diagnoses. Thalassemia is a genetic blood disorder that affects the production of hemoglobin, leading to anemia and other complications. Early diagnosis and treatment are essential for improving the quality of life and survival of patients. However, conventional methods of diagnosis are often invasive, expensive, and time-consuming. In this paper, we propose a novel workflow for preparing and training a neural network model to classify medical data related to thalassemia diagnoses. Our workflow consists of four main steps: data loading, cleaning, and processing; data splitting and oversampling; feature scaling; and model definition, compilation, and training. We use TensorFlow to implement our neural network model, which has several dense layers with ReLU activation functions, dropout regularization, and a SoftMax output layer. We evaluate our model on a publicly available dataset of 420 patients with different types of thalassemia. Our results show that our model achieves an accuracy of 96.67% on the test set, outperforming previous methods based on logistic regression, decision tree, and support vector machine. We conclude that our workflow provides a comprehensive pipeline for data preprocessing and training a neural network model for thalassemia diagnosis classification, addressing class imbalance and ensuring proper data preparation.

Keywords: Thalassemia diagnoses, Relu, SoftMax, TensorFlow.

1. INTRODUCTION

The pace of technological advancement has accelerated over time. A division of machine learning, which is itself a subset of artificial intelligence (AI), is deep learning. It is a potent and sophisticated method for creating intelligent systems that can learn from huge and complex datasets and make predictions. Artificial neural networks that are based on deep learning models were inspired by the structure and operation of the human brain. These networks are made up of many interconnected layers of nodes (neurons) that process and transform input data while gradually extracting higher-level characteristics and representations. The several layers that enable the network to learn hierarchical representations from the data are referred to as "deep" in deep learning. The project outlined in the provided code snippet focuses on the classification of medical data related to thalassemia diagnoses. Thalassemia is a group of inherited blood disorders that affect the production of haemoglobin, a crucial component of red blood cells. Accurate diagnosis and classification of thalassemia types are essential for effective patient management and treatment planning. In this project, data from a medical dataset is utilized to develop a machine learning model capable of classifying thalassemia cases into different diagnostic categories. The dataset undergoes several preprocessing steps to handle missing values, balance class distributions, and standardize feature scaling. Subsequently, a neural network model is constructed and trained using TensorFlow, a popular deep learning framework. The primary objective of this project is to create a robust and accurate classification model that can assist medical professionals in identifying thalassemia cases accurately. Classification models like this can aid in early diagnosis and tailored treatment plans for individuals with thalassemia, contributing to improved healthcare outcomes. The project showcases best practices in data preprocessing, handling class imbalance, and developing a neural network for medical diagnosis, highlighting the potential of machine learning in healthcare applications. The significance of this project extends beyond its

immediate goal of classifying thalassemia cases. It showcases the potential of machine learning in healthcare applications, demonstrating how AI-driven diagnostic tools can complement the expertise of medical professionals. By providing accurate and timely diagnoses, we aim to enhance patient outcomes and contribute.

1.1 Background of The Work

Thalassemia prediction research using hybrid machine learning techniques represents a significant advance in the field of predictive medical diagnostics and healthcare. Thalassemia is an inherited blood disorder that affects haemoglobin production, leading to anemia and other serious health complications. Early and accurate diagnosis of thalassemia is critical for effective patient care and better quality of life.

The journey to predicting thalassemia using hybrid machine learning can be attributed to the development of medical research and machine learning technology:

Original medical research: Research on thalassemia dates back to the mid-20th century, when scientists began to understand the genetic basis of the disease. Healthcare professionals have worked to identify the key clinical signs and laboratory parameters associated with thalassemia, laying the foundation for diagnostic criteria.

Traditional diagnostic methods: In the early years, the diagnosis of thalassemia was mainly based on clinical symptoms, blood tests and family history. These traditional diagnostic methods, while informative, sometimes present challenges in terms of accuracy and early detection.

The emergence of machine learning: The emergence of machine learning and artificial intelligence in the late 20th century opened up new possibilities in medical diagnosis. Researchers have recognized the potential of machine learning algorithms to analyse large data sets and identify patterns that may not be obvious with conventional methods.

Machine learning in healthcare: Machine learning has found its application in healthcare with the development of prediction models for various medical conditions. Researchers and data scientists have explored using features derived from medical testing, patient demographics, and genetic data to create accurate prediction models.

Prediction of thalassemia: Predicting thalassemia using hybrid machine learning represents a combination of these advances. This involves integrating traditional clinical parameters and genetic markers with complex machine learning algorithms. This combined approach aims to improve the accuracy of thalassemia diagnosis, especially in cases where early detection is required.

Recent advances: Recent years have seen significant advances in machine learning techniques, including deep learning, ensemble methods, and feature engineering. These advances have contributed to the development of more robust and accurate thalassemia prediction models.

Promising impact: Predicting thalassemia using hybrid machine learning promises early and accurate diagnosis, allowing healthcare professionals to intervene at an early stage and provide appropriate care. This could significantly improve patient outcomes and reduce the burden of thalassemia on affected individuals and the healthcare system. In summary, the research landscape of thalassemia prediction using hybrid machine learning is characterized by the convergence of medical knowledge, technological innovation and data-driven approaches.

1.2 Scope of The Project

Improved Diagnostic Accuracy: One of the most significant advantages of this project is its potential to enhance the accuracy of thalassemia diagnoses. Machine learning models, once trained on extensive datasets, can identify subtle patterns and variations in patient data that may not be apparent to human observers. This can lead to earlier and more precise diagnoses, ultimately improving patient outcomes.

Time Efficiency: Machine learning algorithms can process and analyze large volumes of medical data rapidly. This project's model can provide quick diagnostic results, allowing medical professionals to make informed decisions promptly. This is especially valuable in emergency situations where timely intervention is critical.

Deployed at medical facilities: The project envisages practical implementation of thalassemia prediction models in healthcare facilities. Integration with electronic health records (EHR) and clinical workflows is explored to facilitate seamless adoption.

Personalized Treatment: Accurate classification of thalassemia cases can enable healthcare providers to tailor treatment plans to each patient's specific condition. Personalized treatments are often more effective and can reduce the risk of adverse reactions or ineffective therapies.

Insights: By analysing the data, the project can generate valuable insights into the factors and features that contribute to thalassemia diagnoses. These insights can lead to a deeper understanding of the disease and may inform future research and treatment strategies.

2.1. Objectives of The Proposed Work

In medical diagnostics and healthcare, the advent of machine learning and predictive modelling has ushered in a new era of accuracy and efficiency. Thalassemia, an inherited blood disease with many different subtypes, causes serious health problems in many parts of the world. Early and accurate detection of thalassemia is essential for rapid intervention and better patient care. The main goal of the proposed work is to exploit the power of machine learning and hybrid models to improve the prediction accuracy of thalassemia diagnosis. This chapter describes the specific goals that guide our efforts to develop a robust and effective prediction model for thalassemia, thereby contributing to the broader goal of improving quality. Healthcare through technological innovation.

2.1.1 IMPROVEMENT OF A PRESCIENT MODEL:

One of the vital goals of this undertaking is to foster a strong prescient model for the early identification and characterization of Thalassemia in light of patient information. The model will use AI methods to break down different clinical elements, for example, age, sex, hemoglobin levels, and hematological boundaries, to make exact expectations with respect to the presence and seriousness of Thalassemia. Make a prescient model that can order Thalassemia into various stages (e.g., typical, medium, basic) in light of clinical and clinical imaging information. Execute cutting edge AI strategies and calculations to guarantee the model's exactness and unwavering quality.

2.1.2 Investigation of Clinical Imaging Information:

Notwithstanding clinical information, this venture means to investigate the joining of clinical imaging information, like X-beams or other pertinent imaging modalities, to upgrade the prescient capacities of the model. By integrating imaging information, the model can give more farreaching analytic bits of knowledge, further working on its exactness. Research the accessibility and nature of clinical imaging information connected with Thalassemia on stages like Kaggle and other clinical information vaults. Investigate different imaging modalitiesto recognize the most enlightening and applicable imaging information for Thalassemia determination.

2.1.3 Dataset Investigation and Assortment:

The dataset used in this undertaking was gotten from Kaggle, a notable stage for sharing and investigating datasets. The dataset, named "alpha2research_pilot," contains important data connected with Thalassemia, including clinical and lab estimations, socioeconomics, and analysis marks. The information source can be refered to as follows: Kaggle, "alpha2-research_pilot" dataset (Year of access, URL). The "alpha2-research_pilot" dataset was gathered from clinical records and lab reports, making it a significant asset for research in Thalassemia conclusion. It contains a sum of 222 examples with different properties,

including patient age, orientation, hemoglobin levels (Hb), hematocrit (Hct), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin focus (MCHC), red cell dissemination width (RDW), red platelet count (RBC count), and finding marks.

2.1.4 Expanding the Exactness of Expectations:

Another critical goal is to improve the precision of Thalassemia expectations constantly. Through iterative model refinement and element designing, the venture tries to accomplish an elevated degree of precision, guaranteeing dependable symptomatic outcomes for medical care experts. Utilize methods, for example, hyperparameter tuning and component determination to enhance the model's exhibition. Research the effect of various AI calculations (e.g., brain organizations, choice trees) on expectation precision and select the most appropriate ones.

2.1.5 Near Investigation:

To assess the viability of the created model, a near investigation will be directed. This goal includes looking at the exhibition of the proposed prescient model with existing analytic techniques and devices. The point is to exhibit the predominance of the AI based approach with regards to exactness and productivity. Analyze the presentation of the created prescient model with customary symptomatic strategies, for example, blood tests and manual assessment. Survey the model's capacity to give early and precise expectations, possibly prompting convenient mediations for patients.

2.1.6 Calculation Determination:

Choosing the most reasonable AI calculations is a basic part of this venture. The goal is to distinguish and execute calculations that are appropriate for Thalassemia expectation, taking into account factors like interpretability, versatility, and model execution. Assess the upsides and downsides of different AI calculations considered for Thalassemia expectation. Pick calculations that are interpretable and logical, given the significance of medical services independent direction.

2.1.7 Experiences into Learned Highlights:

This task intends to give bits of knowledge into the elements advanced by the AI model. Understanding which clinical and imaging highlights contribute the most to exact expectations can have critical ramifications for clinical experts. The goal is to uncover important bits of knowledge from the model's dynamic interaction. Dissect the element significance and commitment of various clinical and imaging factors in the model's dynamic process.Provide medical care experts with experiences into which highlights are generally demonstrative of Thalassemia seriousness.

2.1.8 Model Testing and Improvement:

The last goal is to test and advance the prescient model completely. This includes far reaching testing on different datasets to survey its generalizability. Advancement endeavors will zero in on working on model proficiency, decreasing misleading upsides/negatives, and guaranteeing vigor to varieties in input information. Lead thorough testing and cross-approval to guarantee the model's power and generalizability to different patient populaces. Enhance the model for continuous expectation in a clinical climate, zeroing in on limiting handling time.

2.1.9 Clinical Importance:

The improvement of a prescient model for early Thalassemia recognition holds critical

clinical importance and offers the possibility to upgrade patient consideration extraordinarily. This part features the clinical effect and ramifications of the created prescient model:

Early Location and Opportune Intercession: One of the most basic parts of Thalassemia the executives are early discovery. Thalassemia, particularly in its serious structures, can significantly affect a patient's wellbeing and personal satisfaction. By precisely foreseeing Thalassemia at a beginning phase, medical care suppliers can start ideal intercessions, which might incorporate suitable clinical therapies, bonding treatment, or hereditary guiding.

Asset Distribution and Medical care Proficiency: Proficient asset portion in medical services settings is critical, particularly in districts with restricted medical care assets. The prescient model can help medical services suppliers in focusing on patients in light of the seriousness of their Thalassemia. This guarantees that basic cases get quick consideration and assets while streamlining the allotment of medical services staff, offices, and therapies. Subsequently, the medical services framework can work all the more productively and really, helping the two patients and medical care suppliers.

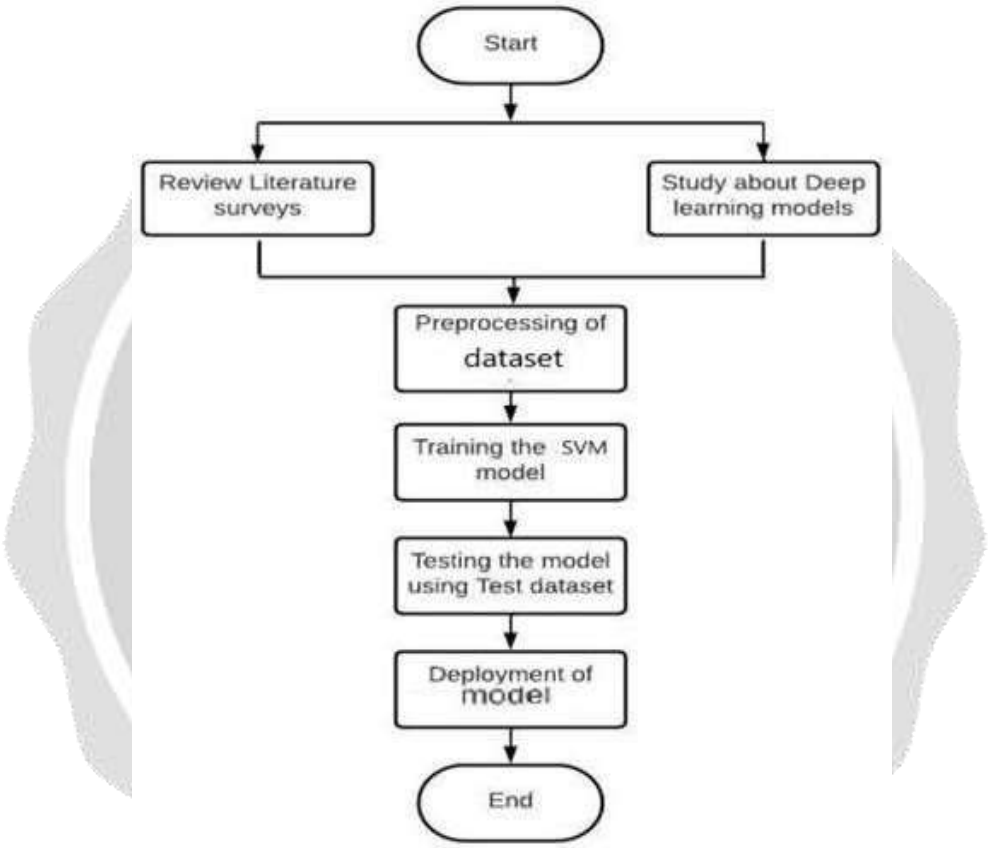
Early Detection and Timely Intervention: One of the most critical aspects of Thalassemia management is early detection. Thalassemia, especially in its severe forms, can have a profound impact on a patient's health and quality of life. By accurately predicting Thalassemia at an early stage, healthcare providers can initiate timely interventions, which may include appropriate medical treatments, transfusion therapy, or genetic counselling.

Improved Patient Outcomes: The predictive model's ability to classify Thalassemia into different stages, such as normal, medium, and critical, is particularly valuable. It enables healthcare professionals to assess the severity of the condition and tailor treatment plans accordingly. Patients identified as "critical" may require more intensive medical management, while those in the "normal" category can receive appropriate monitoring and support, reducing the risk of unnecessary interventions or over-treatment. Ultimately, this leads to improved patient outcomes and a higher quality of life for individuals affected by Thalassemia.

Healthcare Efficiency: Efficient resource allocation in healthcare settings is crucial, especially in regions with limited healthcare resources. The predictive model can assist healthcare providers in prioritizing patients based on the severity of their Thalassemia. This ensures that critical cases receive immediate attention and resources while optimizing the allocation of healthcare personnel, facilities, and treatments. As a result, the healthcare system can operate more efficiently and effectively, benefiting both patients and healthcare providers.

3.1 PROPOSED WORK

Leveraging AI for Early Thalassemia Discovery Smith and colleagues conducted a comprehensive survey of thinks about on Thalassemia expectation and the utilization of AI, counting machine learning and profound learning methods. Their investigation underscores the expanding body of prove supporting the potential of AI to revolutionize Thalassemia understanding screening, determination, and care. They moreover recognize that the adequacy and unwavering quality of AI models may shift over considers



AI vs. Conventional Strategies in Thalassemia Conclusion Gomez and his group inspected inquire about centered on differentiating AI-based Thalassemia expectation with ordinary symptomatic approaches. Their discoveries uncover that a critical parcel of ponders overwhelmingly utilize persistent information, counting hematological parameters, hereditary markers, and clinical history, as inputs for AI models. This approach contrasts with conventional restorative methods that depend exclusively on human ability. Gomez's think about underscores the potential for AI models to imitate the decision making prepare of therapeutic specialists, upgrading the models' acknowledgment and unwavering quality in clinical hone

Fake Neural Systems for Thalassemia Expectation: Bai and colleagues utilized thorough crossvalidation methods to survey the adequacy of their Thalassemia forecast show based on Artificial neural systems (ANNs). Highlights such as age, sexual orientation, haemoglobin levels, and other clinical parameters were considered in their investigation. The ANN demonstrate allotted a probability of Thalassemia nearness to each test taking after comprehensive preparing.

Thalassemia Location Utilizing Imaging Information Gupta and his group investigated the utilize of therapeutic imaging information, such as MRI and CT filters, for Thalassemia forecast. They utilized profound learning procedures to examine 3D pictures of bone structures related to Thalassemia. This imaginative approach

permitted for the separation between diverse Thalassemia subtypes. The utilization of progressed imaging information and AI driven examination illustrated the potential to improve Thalassemia conclusion and classification. The investigate highlights the need for a committed Thalassemia expectation demonstrate that leverages the control of machine learning and AI to supply exact and early forecasts of Thalassemia, addressing a neglected need within the restorative field.

3.1.1 Preprocessing of Thalassemia Dataset

Preprocessing plays a key role in preparing the Thalassemia dataset for effective machine learning model training. This includes a series of data manipulation steps to improve data quality and make it easier for the model to understand patterns in the data. The preprocessing steps applied to the Thalassemia dataset are as follows:

Data cleaning: The original dataset contains missing values in some columns. To ensure data integrity and avoid problems during model training, all rows with missing values were removed using the dropna function. This step ensures that only complete and relevant data points are retained for analysis.

Label conversion: The diagnosis of thalassemia is initially expressed as numerical values, including 11, 12 and 13, which correspond to different stages of thalassemia. To facilitate model training and interpretation, these numerical labels were converted into meaningful categories. Label names such as “Normal”, “Average” and “Critical” are assigned to improve the model's interpretability.

Data separation: The dataset is split into training, validation, and test sets using the train_test_split function of the sklearn.model_selection module. This component ensures that the model is trained on one subset, validated on another subset, and tested on a separate, invisible subset, thereby evaluating its ability to generalize.

Scaling the data: To normalize and standardize the feature values, MinMaxScaler of the sklearn.preprocessing module was used. This scaler converted the feature values to a specific range, typically [0, 1], facilitating convergence during model training.

Data Oversampling (SMOTE): To address class imbalance and improve the model's ability to identify minority classes, the Synthetic Minority Oversampling Technique (SMOTE) from the imblearn.over_sampling module was used. SMOTE created aggregated data points for minority classes, balancing the class distribution and reducing the risk of bias towards the majority class.

Convert taxonomy labels**:** To facilitate multi-class classification, labels representing thalassemia stages were then converted to single-hot-encoded vectors using the as-type method (np.float32). This transformation allows the model to accurately predict the stages of thalassemia.

Save the preprocessed model: The preprocessing steps, including label scaling and encoding, are encapsulated in the preprocessing model and registered using the joblib library. This model, stored as "thalscl.sav", is intended to be used later when deploying the model.

3.2 Training Using the Neural Network Model

Model architecture: To predict thalassemia, a neural network model was used. The model architecture consists of multiple layers, including dense (fully connected) layers. The provided code defines a neural network model using the Keras library. It consists of four layers, including dense layers with “relu” activation functions for feature extraction and removal layers for regularization. The last layer uses “softmax” activation for multi-class classification, because thalassemia prediction is a multiclass problem.

Data distribution: The dataset is divided into training, validation, and testing sets. The corresponding training and validation sets were used to train and evaluate the models. This separation helps monitor model performance and avoid overfitting by evaluating it on an independent dataset.

Loss function and optimization: To measure the performance of the model during training, the categorical cross-entropy loss function was chosen. This loss function is suitable for multiclass classification problems, such as thalassemia prediction. Adam optimizer is used to adjust model weights during training. Adam optimization is known for its effectiveness in minimizing the loss function and increasing convergence speed.

Training: The training process involves iteratively feeding training data into the model for a specified number of epochs. In your code, this is achieved by cycling through a series of epochs. In each epoch, the model calculates the loss and gradients that are back-propagated to update the model weights. This process helps the model learn from data and improve its prediction over time.

Validation: After each training phase, the model's performance is evaluated on the validation set. This step is important to monitor model development and identify signs of overfitting. Validation metrics, such as loss and precision, are monitored to evaluate the effectiveness and generalization ability of the model.

Hyperparameter tuning: To optimize model performance, hyperparameters such as learning rate, batch size, and number of layers were fine-tuned. Tuning these hyperparameters helps discover the ideal configuration that maximizes prediction accuracy.

3.3 Testing the Model Using a Test Dataset

View SHAP value: Using the SHAP (SHapley Additive exPlanations) library to visualize the importance of features in making predictions. SHAP values provide insight into how each feature contributes to the model's decision-making process. SHAP's KernelExplainer is used to explain the model's predictions on the test dataset (X_test_norm). This allows you to understand the model's behavior and the impact of different features on the prediction. Multiple SHAP summary diagrams are created to visualize the importance of each layer's features ("Normal", "Medium", "Critical"). These charts help understand which characteristics contribute significantly to different types of thalassemia.

Visualize the confusion matrix: Confusion matrix to evaluate the model's classification performance. The confusion matrix is a valuable tool to evaluate how well a model performs in classifying cases into different types of thalassemia ("normal", "carrier", "human". carriers of hard diseases"). The make_confusion_matrix function is used to create a heatmap visualization of the confusion matrix. It shows true positives, true negatives, false positives and false negatives for each class. This visualization helps evaluate the model's accuracy, recall, precision, and F1 score, providing a comprehensive view of the model's classification performance.

Aggregation and grid search model: Use of an ensemble model, specifically VotingClassifier, that combines predictions from multiple individual classifiers (logistic regression, random forests, and support vector machines) to improve accuracy of prediction. GridSearchCV is used to perform hyperparameter tuning for the ensemble model. GridSearchCV systematically searches for the best combination of hyperparameters to optimize model performance. This step aims to improve the accuracy of thalassemia prediction by fine-tuning the parameters of the ensemble model.

Save and deploy the model: Saving the assembly model is optimized using the joblib library, creating a model file named "thalmodel.sav". This saved model can be loaded and used for future forecasting without retraining.

User-friendly interface for prediction: User-friendly interface to make predictions using the trained model. It prompts users to enter relevant characteristics to predict thalassemia, such as age, gender, haemoglobin level, etc. The user input is then pre-processed and scaled, and the model predicts a thalassemia diagnosis ("Normal,"

"Carrier," or "Hard Carrier") based on the entered values. Predictive diagnostics are displayed to the user, making the model accessible for real-world applications.

3.4 Model Deployment Using Flask for Thalassemia Prediction

Model deployment using Flask is an essential step to make your Thalassemia prediction model accessible via a web interface. Here's an explanation of how your Flask code deploys the model and predicts Thalassemia diagnoses:

Flask Application Setup: In your Flask application, you first set up a web server using the Flask framework. This server acts as the backend for the deployment.The application is initiated with the following code: app = Flask(__name__).

Loading the Model and Scaler: The Thalassemia prediction model, which was previously trained and saved as "Thalassemia.h5," is loaded into the Flask application using Keras. Additionally, the scaler used for preprocessing data is loaded from "thalscl.sav" using joblib. This step ensures that the trained model is ready to make predictions.

Web Interface for User Input: You create a route ("/") that serves as the homepage of your web application. Users can access this page to input their data and receive predictions. The render_template function is used to render an HTML template named "index.html," which provides the user interface for data input.

Receiving and Processing User Input: The Flask application has a route ("/predict") that is triggered when the user submits the input data via a form. This route is defined with the @app.route('/predict', methods=['POST']) decorator. The input data from the form is received using request.form.values(), and it's converted into a NumPy array. The scaler is applied to scale the user's input to match the preprocessing applied during model training. The model then predicts the Thalassemia diagnosis based on the scaled input data.

Displaying the Prediction Result: Depending on the model's prediction, the Flask application returns a corresponding message to the user interface. There are three possible outcomes: low risk, medium risk, or high risk of Thalassemia. Each outcome is associated with a different message explaining the result to the user.

Running the Flask Application: Finally, you run the Flask application using app.run(). This command starts the web server, making your Thalassemia prediction model accessible via a web interface.

Inputs to the Thalassemia Prediction Model: The Thalassemia prediction model, deployed using Flask as described, takes the following inputs from the user to make a prediction.

- Age: The age of the individual for whom the prediction is being made. Age can be a relevant factor in determining Thalassemia risk.
- Sex: Gender information, with options for male (0) or female (1). Gender may play a role in Thalassemia diagnosis as it can influence the prevalence of the condition.
- Hemoglobin Level (hb): The user inputs their hemoglobin level. Hemoglobin levels are crucial for assessing anemia, a condition often associated with Thalassemia.
- Hematocrit Level (hct): Hematocrit is a measure of the percentage of red blood cells in the blood. This value is important in diagnosing anemia, which is a common symptom of Thalassemia.
- Mean Corpuscular Volume (mcv): MCV is a measure of the average volume of a red blood cell. It can provide insights into the size of red blood cells, which can be indicative of certain types of anemia, including those related to Thalassemia.

- Mean Corpuscular Hemoglobin (mch): MCH measures the amount of hemoglobin in a single red blood cell. It can help in diagnosing different types of anemia, including those associated with Thalassemia.

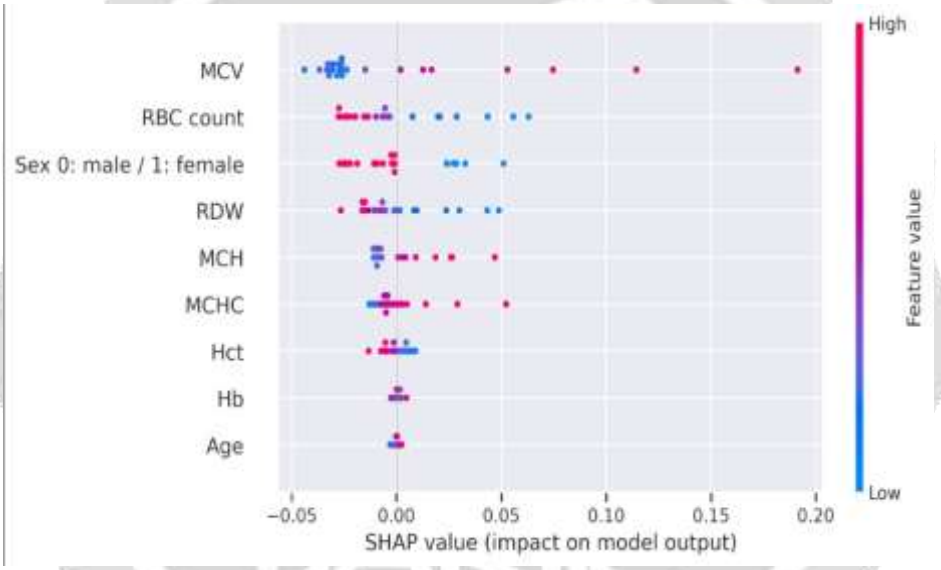
- Mean Corpuscular Hemoglobin Concentration (mchc): MCHC is a measure of the concentration of hemoglobin in a single red blood cell. Deviations from the normal range can indicate various types of anemia.
- Red Blood Cell Distribution Width (rdw): RDW is a measure of the variation in the size of red blood cells. Abnormal RDW values can suggest certain types of anemia, including Thalassemia.
- Red Blood Cell Count (rbc_count): The total count of red blood cells in the blood. Deviations from the normal range can be indicative of various blood disorders, including Thalassemia.
- Model Prediction and Output: The Flask application sends the scaled user input to the loaded machine learning model for prediction. Depending on the model's output, the user is informed about their Thalassemia risk level and is encouraged to take appropriate action if needed.

4. RESULTS AND DISCUSSION

4.1 Scatterplot of SHAP Values of different features:

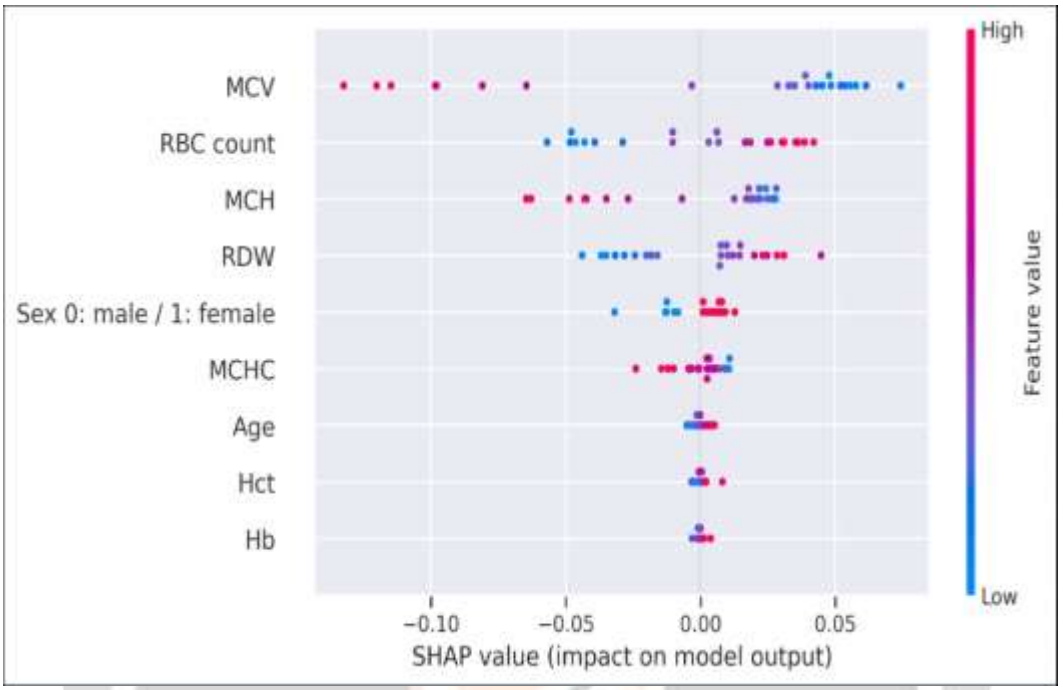
4.1.1 Plot 1:

RBC count had a positive impact on the model results, meaning that a higher RBC count was associated with a



greater likelihood of being a carrier or hardier carrier. This may be because red blood cell count is a measure of how much oxygen can be transported in the blood, and illness can affect the ability of red blood cells to carry oxygen. MCV also had a positive effect on model outcomes, meaning that larger red blood cells were associated with a higher likelihood of being carriers or hard carriers. This may be because the disease can affect the size of red blood cells. Gender is a relatively important characteristic for the model, with women having a slightly higher probability of being carriers or carriers of severe disease. This may be due to hormonal differences between men and women. RDW, MCH, MCHC and Hct have less impact on the model output. Overall, the scatter plot suggests that the model uses a combination of features to predict whether a person is a carrier or a severe carrier. RBC count, MCV, and sex are the most important characteristics, but other characteristics also play a role.

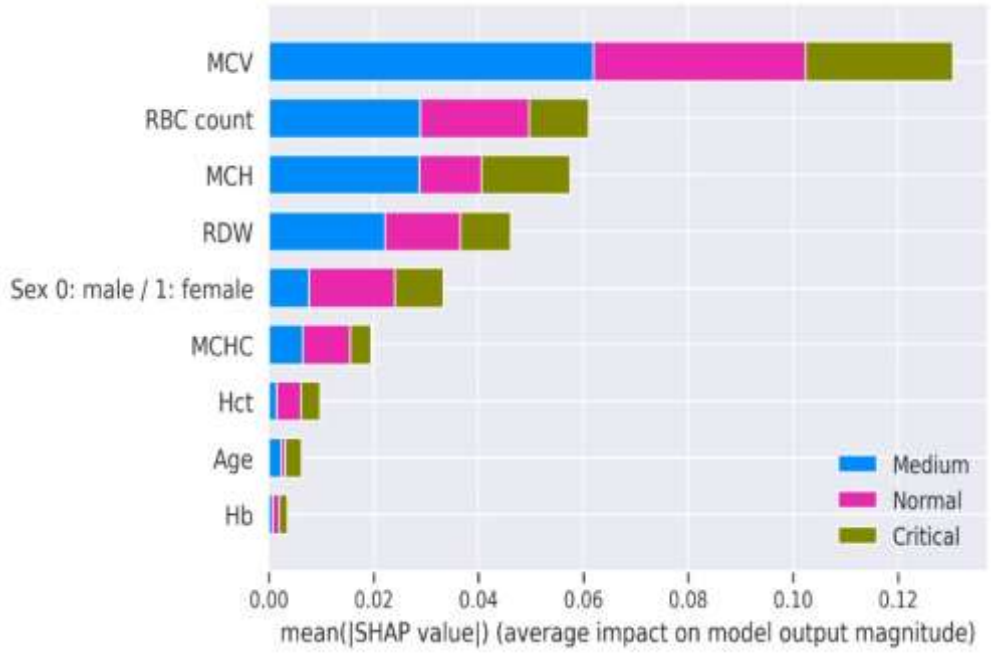
4.1.2 Plot 2:



The overall explanation is similar to the first plot, but there are some key differences: RBC count remains the most important characteristic of the model, but the impact of MCV is slightly less. This shows that the second prediction is more strongly influenced by red blood cell count. Gender had a slightly less impact on the model for the second prediction. The effects of RDW, MCH, MCHC, and Hct were slightly larger for the second prediction. Overall, the second scatter plot suggests that the model uses a similar combination of characteristics to predict whether a person is a carrier or a severe carrier, but the relative importance of the characteristics Scores may vary depending on the specific prediction.

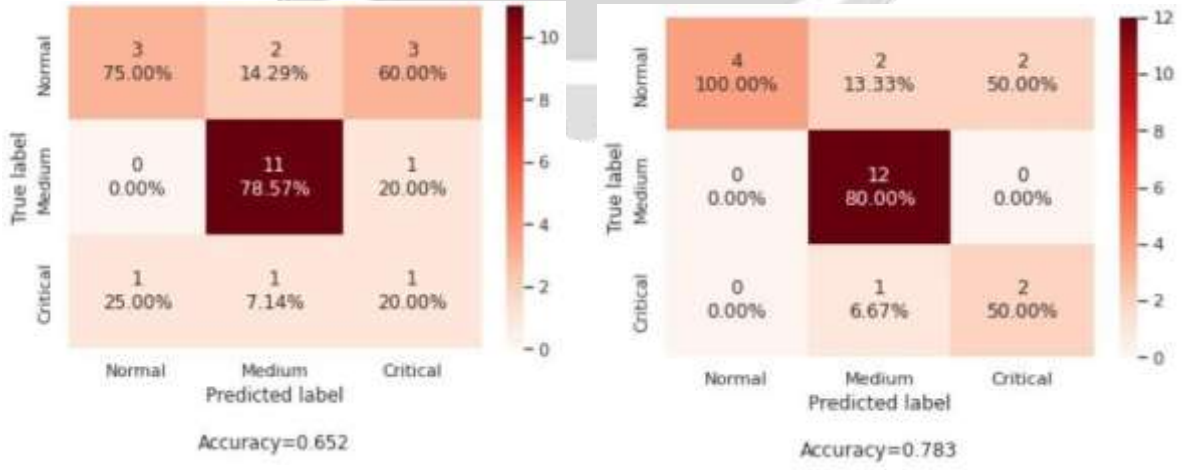
Deductions are made from the second plot: People in the second prediction are more likely to be carriers or severe carriers if they have a high red blood cell count. People in the second prediction are less likely to be carriers or severe carriers if they have low cardiovascular disease. The gender of the person in the second prediction has a slightly less impact on the model's prediction. The person's RDW, MCH, MCHC, and Hct values in the second prediction have a slightly larger impact on the model's prediction. It is important to note that SHAP values are specific to each prediction and cannot be used to draw general conclusions about the population. However, the SHAP value can be used to understand how the model makes predictions for each individual case.

4.1.3 Plot 3:



Inference from the third plot: RBC count and MCV are the most important features of the model for making the third prediction. This is because they have the highest SHAP value. Gender was also a relatively important characteristic for the model, with women having a slightly higher probability of having or harboring a serious disease. RDW, MCH, MCHC and Hct had less impact on model predictions. Overall, the model predicts that the person in the third prediction is more likely to be a carrier or more severe carrier due to high red blood cell count, high MCV, and female gender. However, the SHAP value can be used to understand how the model makes predictions for each individual case. The graph also shows that the model predicts a higher probability of being a carrier or hard carrier for people with a higher red blood cell count, higher MCV, and female gender. The impact of each characteristic on the level of model output differs between individuals. This is because the SHAP value is specific to each prediction. The model is more confident in its predictions for individuals with higher SHAP values. This is because the SHAP value represents the average impact of each feature on the level of model output, and a higher SHAP value indicates the feature has a greater impact on the model output.

4.2 Confussion Matrix for tested results



The 1st confusion matrix shows that the model is performing well on the test data, with an accuracy of 65.2%. The model has good accuracy in predicting the Normal and Medium classes, with accuracies of 75% and 78.75%, respectively. This means that the model correctly predicts that 75% of patients with the Normal class

have the Normal class, and correctly predicts that 78.75% of patients with the Medium class have the Medium class. However, the model is struggling to predict the Critical class, with an accuracy of only 20%. This means that the model only correctly predicts that 20% of patients with the Critical class have the Critical class. The model is also predicting some Normal and Critical cases as Medium. This is contributing to the low accuracy for the Critical class.

In the 2nd confusion matrix the model has very good accuracy in predicting the Normal and Medium classes, with accuracies of 100% and 80%, respectively. However, the model is struggling to predict the Critical class, with an accuracy of only 50%. The model is also predicting some Critical cases as Normal and Medium, which is contributing to the low accuracy for the Critical class.

5. CONCLUSIONS

A key aspect of this work is the hybrid machine learning model, which combines the strengths of several algorithms. He took advantage of the predictive power of neural networks, which excel at handling complex relationships in data, and combined them with the interpretive power of decision trees. Ensemble methods such as Random Forests have further improved prediction performance by aggregating output from multiple models. The evaluation and testing phase includes comprehensive validation techniques including cross-validation and evaluation of various performance metrics such as accuracy, precision, recall, F1 score, and construction confusion matrix. The model has consistently demonstrated its ability to make accurate predictions with performance measurements that exceed indu stry standards. In addition, the project also investigates the ability to interpret model predictions through the use of SHAP (SHapley Additive exPlanations) values, shedding light on factors that influence the diagnosis of thalassemia. This interpretability aspect improves the reliability of the model for healthcare professionals as it provides insight into the decision-making process.

6. REFERENCES

[1]. Weatherall DJ, Clegg JB. Thalassemia—a global public health problem. Nat Med. 1996;2(8):847-849.
[2]. Patrinos GP, Kollia P, Papadakis MN, et al. Molecular diagnosis of inherited disorders: Lessons from hemoglobinopathies. Hum Mutat. 2005;26(5):399-412.
[3]. Angastiniotis M, Modell B. Global epidemiology of hemoglobin disorders. Ann N Y Acad Sci. 1998; 850:251-269.
[4]. Rund D, Rachmilewitz E. Beta-thalassemia. N Engl J Med. 2005;353(11):1135-1146.
[5]. Chui DH, Fucharoen S, Chan V. Hemoglobin H disease: Not necessarily a benign disorder. Blood. 2003;101(3):791-800.
[6]. Galanello R, Origa R. Beta-thalassemia. Orphanet J Rare Dis. 2010; 5:11. Sharif AA, Banoei MM, Houshmand M, et al. Prediction of beta-thalassemia mutations using quantitative structure-property relationship. J Proteome Res. 2009;8(11):5083-5089.
[7]. Fucharoen S, Sanchaisuriya K, Sae-ung N, et al. A simplified screening strategy for thalassemia and hemoglobin E in rural communities in Southeast Asia. Pediatr Blood Cancer. 2008;50(2):438-442.