PREDICTION ON THALASSEMIA CARRIER USING MACHINE LEARNING CLASSIFICATION

Kashfia Hasan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
kashfia2812@gmail.com

***Abstract*—One genetic illness, referred to as a blood disorder that affects the generation of hemoglobin, is thalassemia. The loss in hemoglobin caused by this genetic condition makes it impossible for adequate oxygen to reach all of the body's tissues. Anemia, which frequently results in light skin pigmentation, weariness, weakness, and more severe consequences, was also a problem for people or patients. While identifying disorders is crucial to giving patients the right medical care.The goal of this project is to identify methods for diagnosing the condition by looking for patterns in the data through classification analysis and applying machine learning techniques for early prediction in patients with Alpha Thalassemia based on their physical examination. Six classifiers were applied to the Alpha Thalassemia dataset in this project, which was obtained from the Kaggle Dataset repository. Finally, this study can identify which classifier performs better in terms of accuracy and categorize the patients as alpha thalassemia carriers or normal based on the different performance parameters.**

I. Introduction

Relevant scientific fields have accumulated a vast amount of biological knowledge during the past few decades. The creation of effective and efficient computing systems for storing, analyzing, and interpreting this data has been made necessary by the deluge of such information in the form of genomes, protein sequences, data on gene expression, and so on. Then comes bioinformatics, a branch of applied computing science that uses a variety of methodologies from chemistry, biochemistry, artificial intelligence, computing and informatics, statistics, and mathematics to tackle biological problems, particularly at the molecular level.((Cios et al., 2005, Kelemen et al., 2008, Gusfield, 2004, Valentini et al., 2009, Smolinski et al., 2008a, Smolinski et al., 2008b). Sequence alignment, gene discovery, genome assembly, protein structure alignment, protein structure prediction, gene expression prediction, protein-protein interaction prediction, and evolution modeling are some of the major research initiatives in this area. hence, In other words, the use of computer techniques to create biological discoveries is what is meant by the term "bioinformatics." s (Anon., 2013, Smolinski et al., 2008a, Smolinski et al., 2008b)

II. Related Work

A. Literature Review

A deficiency of hemoglobin, the blood protein in charge of carrying oxygen to the tissues, characterizes a group of blood diseases known as thalassemia. Thalassemia, which means "sea blood" in Greek, gained its name because it was first found in people who lived close to the Mediterranean Sea, where it is frequent. Despite being more prevalent among those with ancestors from the Mediterranean, Middle East, and southern Asia, thalassemia genes are distributed throughout the world. (Editors, 2020) Native Americans and certain northern Europeans have also been found to have thalassemia. Those with African ancestry tend to have a noticeably milder condition. It is suggested that the potentially fatal thalassemia gene is retained in some populations because it provides some malaria protection in the homozygous state.

There are several types of thalassemia that have been identified, including those with one, two, three, or more globin chains, al- though the most common types, such as thalassemia and thalassemia, are the most harmful to human mortality [6]. For a precise medical diagnosis, optimized technology and machine learning methods are used. The suggested system for these research is also meant to examine or watch the accuracy of alpha-thalassemia carriers in contrast to healthy hemoglobin. With the majority of prediction algorithms used in machine learning today, it is possible to avoid and prevent mistakes that a specialist would probably make when diagnosing the test samples produced. In this study, ML techniques were used as a classification technique with the goal of classifying a specific set of unstructured data into cate- gories. The purpose, label, and categories are additional definitions for the classes. A conclusion can be drawn from data input training using categorization predictive modeling. By reading the samples, the classifiers will determine the class to which they belong. [1] proposed and compared the most accurate of four algorithms to distinguish between categories as part of their work on machine learning classifiers

1) K-Nearest Neighbor (kNN): An approach known as k-NN classifies unknown cases by using the nearest already existing examples from the target space as the feature. This method's core tenet is that instances with the same feature are more likely to belong to the same class if they are located close to one another and remain in that class than if they are farther away.(Thirumuruganathan, 2010). The classification technique finds the most frequent class label among an unknown instance's k closest neighbors with the closest distance between the unidentified features and learned features and assigns a label as a class to the instance. It is common practice to measure the distance between instances using the Euclidean distance.(Thirumuruganathan, 2010)

2) Naive Bayes (NB): Naive Bayes classifiers are another type of machine learning algorithm that primarily rely on the Bayes theorem to determine classification results. The Nave Bayes classifier is founded on the idea that a feature's impact on a class is unrelated to its presence in comparison to another feature. To put it simply, the Nave Bayes technique generates good predictions on the basis of asserting that features are independent based on specific constraints, i.e., that the values of no feature for a given class label rely on the values of any other features. [1]

3) Support Vector Machine (SVM): Support Models for supervised machine learning called vector machines can be ap- plied to both regression and classification tasks. SVM can resolve independent and nonseparable issues in both linear and nonlinear domains. The best separating hyper- plane that divides each data point into two classification points is chosen using SVM in order to categorize the data points as correctly as feasible. There are two actions that must be completed in order to generate an SVM classifier. The mapping of the N-dimensional input data into a sizable feature space is finished by a fitting kernel function. The SVM then finds which separating hyperplane in this feature space has the largest separation margin between the data.

4) Dimensionality Reduction: There are numerous reasons why machine learning could be accurately categorised when it comes to classification. These factors, also known as characteristics, are most likely what these causes are. If the number of features is excessively great in compar- ison to the other variables collected, it would be difficult to visualize the training set. There are times when the features are the same and are more likely to be redundant. The purpose of dimensionality reduction was thus clarified. Dimensionality reduc- tion is a technique for reducing the number of pointless characteristics in the data to improve machine learning model prediction and accuracy. The most frequent dimen- sions in dimensionality reduction are those involved in feature selection and feature extraction. In the advancement of machine learning, classification applications make the most use of these two-dimensionality reduction techniques.

[8] investigation is ongoing "Performance evaluations of Tomosyn- thesis breast lesions images were supervisedly categorized using shallow and deeper neural networks "Because of its abilities, the author decided to extract features using

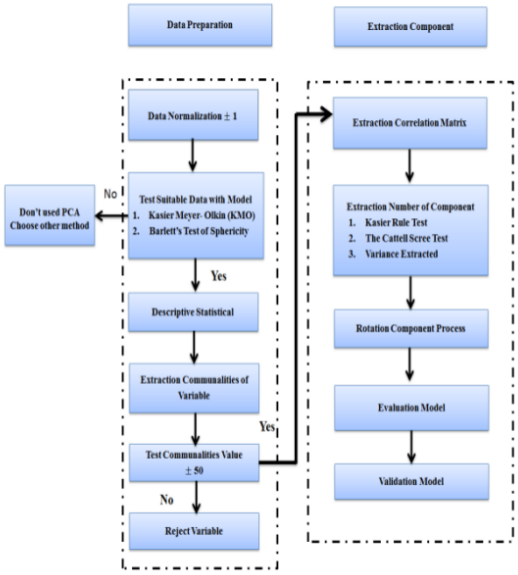

Fig. 1. data preparation step begins with the normalization in [9]

CNN. Above the most effective ANN and non-neural learning methods. According to AlAgha et al. (2018), CNN models were assessed as feature extractors, with each final layer acting as an input for several learners. Activation normalization is useful for generating minor performance increases when training CNNs with learners, but image augmentation is required to significantly improve classification performance, especially when the number of samples per class is uneven (Bevilacqua et al., 2019)

III. System Design

A. Methodology

B. Research Framework

There are three phases involved for comparison the performance among the six classifiers in predicting alpha thalassemia carriers. The goal of this experiment is to ensure that it been conducted on the right way with proper workflow of frameworks. There are three phases included in the workflow, the first phases are more focused on making literature review and studies on methodologies used in the previous studies, next phase would be designing and developing an algorithm and lastly, the result analysis and discussion phases. To make the phases visible, a flowchart has attached as Figure 2, as the flows of this research methodology of this project. The machine learning techniques used in this project are -

- KNN
- Random Forest
- Decision Tree
- Gaussian Naive Bayes
- Support Vector Machines
- Logistic Regression

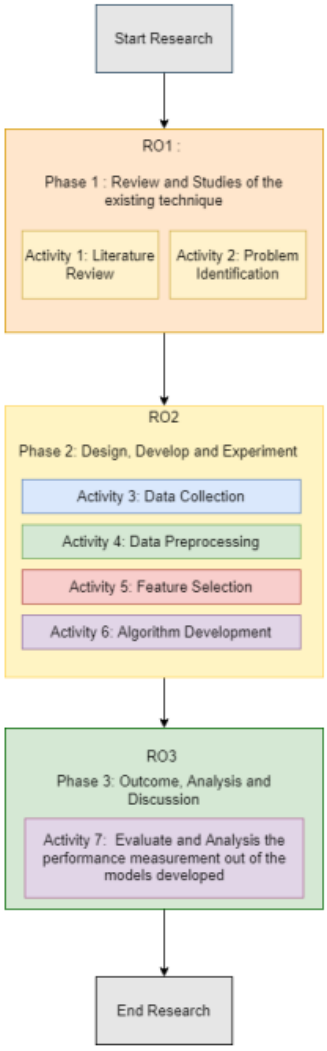

Fig. 2. Flow of the research methodology

C. Design, Developing and Implementation

During this phase, the methods include getting the dataset, pre-processing the data, feature extraction, and model construction. The dataset used in this research is the alpha thalasssemia disease collected from AlphaPred: A Machine Learning Tool to Aid Alpha Thalassemia Screening studies from [7]. The dataset was available on Kaggle dataset and openly access to many users online to perform research on the dataset uploaded on the respiratory of Alpha Thalassemia. To improve the predictive accuracy of the classifier's models, the dataset used should be more exact and accurate. The dataset may have some missing value or irrelevant attributes or more likely a noise contained inside the values of dataset. These problems and challenge can be solved efficiently to obtain the best results for the classification accuracies. The solution from the problem is known as data pre-processing. The activities involved in data pre-processing by examining the missing values and data normalization, Values can be addressed with by either eradicating them or replacing them with an average, frequency, maximum, or minimum. Eradicating the instances might reduce the amount of data and the quality of prediction. Next is data normalization, this is step of data preprocessing since the dataset might contain different scales in a wide range. The process is to uniform the instance values from different ranges into the range 0 to 1 by generating new values while maintaining the general distribution and ratio of the given dataset.

1) Data Preparation: The dataset in this study of Alpha thalassemia classification is obtained from the Kaggle dataset repository and contains samples from 202 samples with 15 columns. This dataset is the records from 288 individuals and belong to Postgraduate institute of Medicine, University of Colombo, Sri Lanka. The dataset consists full blood data count, Hemoglobin variant data via High Performance Liquid Chromatography (HPLC) and alpha thalassemia carrier status obtained via a genetic diagnosis based on the presence of mutations

2) Data Pre-Processing: Data pre-processing is a set of procedures for converting raw data obtained from data extraction into a "clean" and "tidy" dataset. Pre-Processing is the process of evaluating and enhancing the quality of data as it were properly organized which could be fitted to statistical analysis. There are important steps which involved in preprocessing data. Pre-processing data includes the following steps: data cleansing, data integration, data transformation and data reduction.

3) Data Normalization: Another vital step in preprocessing data used since the dataset might contain different scales in a wide range is normalization. The process is to uniform the instance values which are numerical variables from different ranges into the range of 0 to 1 by generating new value while maintaining the general distribution and ratios of the given dataset. On this study approached the MinMaxScaler () method.

D. Result and Analysis

In this section, the result that I got form all the classifiers is described and analyzed. From the experiment the svm model produced the highest accuracy rate for predicting alpha thalassemia, which is 72.13%. The lowest rate was generated by Naive Bayes Classifier which is 34.43%. Although the Logistic Regression and Support Vector Machine shows the same accuracy rate. The other classifiers' result is shown below-

Algorithm	Accuracy
Random Forest Classifier	70.49
Decision tree classifier	52.46
K Nearest Neighbors Classifier	60.66
Naive Bayes	34.43
Logistic Regression	72.13
Support Vector Machines	72.13

TABLE I
ACCURACY SCORE OF DIFFERENT ALGORITHMS

Therefore it is evident that the support vector machine and logistic regression produces the best results and these are the most successful approaches to predict alpha thalassemia among patients.

IV. Conclusion

There are four primary elements that played significant roles in this research: data preparation, data pre-processing, feature selection, and classification. Only Alpha Thalassemia is the subject of the research's dataset. The goal of this study is to provide a more expeditious and effective method for decreasing characteristics, which would aid in the categorization process and speed up patient care.The results found from this experiment is rather accurate for for further accuracy check and research a larger dataset is needed. This is only a comparison between all the classifiers used here so far.

References

[1] AlAgha, Alaa S., et al. "Identifying -Thalassemia Carriers Using a Data Mining Approach: The Case of the Gaza Strip, Palestine." Artificial Intelligence in Medicine, vol. 88, June 2018, pp. 70–83. DOI.org (Crossref), https://doi.org/10.1016/j.artmed.2018.04.009.

[2] Purwar, Shikha, et al. "Classification of Thalassemia Patients Using a Fusion of Deep Image and Clinical Features." 2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence), IEEE, 2021, pp. 410–15. DOI.org (Crossref), https://doi.org/10.1109/Confluence51648.2021.9377054.

[3] Chong, Joana, et al. "Machine-Learning Models for Activity Class Prediction: A Comparative Study of Feature Selection and Classification Algorithms." Gait Posture, vol. 89, Sept. 2021, pp. 45–53. DOI.org (Crossref), https://doi.org/10.1016/j.gaitpost.2021.06.017.

[4] Amendolia, S. R., et al. "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. 1–2, Nov. 2003, pp. 13–20. DOI.org (Crossref), https://doi.org/10.1016/S0169-7439(03)00094-7.

[5] Aszhari, F. R., et al. "Classification of Thalassemia Data Using Random Forest Algorithm." Journal of Physics: Conference Series, vol. 1490, no. 1, Mar. 2020, p. 012050. DOI.org (Crossref), https://doi.org/10.1088/1742-6596/1490/1/012050.

[6] Rooks, Helen, et al. "A Novel 506kb Deletion Causing Thalassemia." Blood Cells, Molecules, and Diseases, vol. 49, no. 3–4, Oct. 2012, pp. 121–27. DOI.org (Crossref), https://doi.org/10.1016/j.bcmd.2012.05.010.

[7] Nival Chamara Kolambage, et al. AlphaPred: A Machine Learning Tool to Aid Alpha Thalassemia Screening. 2022. DOI.org (Datacite), https://doi.org/10.13140/RG.2.2.24970.95686.

[8] Bevilacqua, Vitoantonio, et al. "A Performance Comparison between Shallow and Deeper Neural Networks Supervised Classification of Tomosynthesis Breast Lesions Images." Cognitive Systems Research, vol. 53, Jan. 2019, pp. 3–19. DOI.org (Crossref), https://doi.org/10.1016/j.cogsys.2018.04.011.

[9] Alkrimi, J. A., Tome, S. A., George, L. E. (2019). Classification of Red Blood Cells using Principal Component Analysis Technique. European Journal of Engineering Research and Science, 4(2), 17–22. https://doi.org/10.24018/ejers.2019.4.2.1007