DOI: https://doi.org/10.4038/sljas.v25i3.8145

Sri Lankan Journal of Applied Statistics, Vol (25-3)

Detection of β - Thalassemia Carriers using Data Mining Techniques

G. K. Subasinghe[1,*], N.V. Chandrasekara[1], and A. P. Premawardhena[2]

[1]Department of Statistics & Computer Science, Faculty of Science, University of Kelaniya, Sri Lanka

[2]Department of Medicine, Faculty of Medicine, University of Kelaniya, Sri Lanka

*Corresponding author: gayathri.subasinghe97@gmail.com

Received: 3rd October 2024/ Revised: 7th November 2024/ Published: 30th December 2024

ABSTRACT

Thalassemia, a genetic blood disorder, presents a significant challenge in Sri Lanka due to its high prevalence. Traditional methods of identifying thalassemia carriers, such as genetic and blood testing, are both costly and time-consuming, and potentially not available for certain demographic groups. However, there haven't been many studies done on the efficacy of data mining models for thalassemia carrier detection, therefore the field is still in its infancy. As such, evaluating their accuracy and utility in clinical practice is crucial. This study aims to develop a time-efficient model to detect the β-thalassemia carriers, which can reduce the time to take a decision and develop the built model as a decision support tool. Eight blood parameters - including RBC, HGB, HCT, MCV, MCH, MCHC, RDW, and HbA2 were selected based on literature. Two model-fitting approaches were introduced, each under different data selection methods: Method 1: Model fitting before handling the class imbalance problem and Method 02: Model fitting with random oversampling technique. Support Vector Machine (SVM) and Probabilistic Neural Network (PNN) models were utilized for β-thalassemia carrier detection. Method 2 exhibited superior performance, especially with the PNN Model 2, achieving an impressive 98.75% overall classification accuracy. Moreover, the implemented PNN Model 2 could be utilized as an efficient decision-support tool, offering both time and cost savings in identifying β-thalassemia carriers. Nonetheless, for further investigation, consulting a medical expert

is recommended.

Keywords: Class-imbalance, Support Vector Machine, Probabilistic Neural Network, β- thalassemia carriers

1 Introduction

The production of haemoglobin, the protein in red blood cells that carries oxygen throughout the body, is impacted by a genetic blood condition known as thalassemia. This condition is caused by mutations in the genes that control the production of haemoglobin, which results in a deficiency in the amount or quality of haemoglobin produced. Thalassemia is a global health problem, particularly in the Mediterranean, Middle East, and Southeast Asia, where it is more prevalent (Mohammed and Al-Tuwaijari, 2021). In most severe cases, the illness can be fatal, hence it must be managed by identifying those who carry the thalassemia genes. A person who carries one or more of the faulty genes causing thalassemia but does not have the disease itself is called a thalassemia carrier, or they have the thalassemia trait (NHS, 2022).

Three categories are used to categorise thalassemia disease: Thalassemia Major (Thal-M), Thalassemia Intermediate (Thal-I), and Thalassemia Trait or Minor (Thal-T). While Thal-T patients typically do not require any specific therapy, those with Thal-M and Thal-I must receive frequent blood transfusions throughout their lives in addition to other treatments. Anaemia is what causes the symptoms of thalassemia. Anaemia symptoms include headaches, paleness, shortness of breath, dizziness and fainting, and a fatigued or weak feeling. People may have mild anaemia or no symptoms at all, depending on the type of thalassemia they have. First, the Complete Blood Count (CBC) test is used to determine whether thalassemia is present. Blood samples from suspicious individuals are then submitted for testing using either Capillary Electrophoresis (CE) or High Performances Liquid Chromatography (HPLC) if the red cell indices/Hb are still low (Elshami and Alhalees, 2012).

Those who have moderate (intermedia) or severe (major) forms of thalassemia typically find out about their condition in childhood since they have symptoms of severe anaemia early in life. People with less severe (minor) variants of thalassemia might not be aware of their condition until they exhibit anaemia symptoms, or until a regular blood test results in anaemia, or until a test is done for another reason (CDC, 2023). The age of presentation and transfusion history were used to distinguish between thalassemia major and intermedia. Patients were categorised as having thalassemia major if they required more than eight transfusions per year, and as having thalassemia intermedia if they required fewer (Beamish et al., 2016).

Thalassemia is always inherited by children from their parents; it can also arise as a result of a gene mutation. Typically, a kid does not show symptoms until the thalassemia gene is present in both parents. If the child received the

thalassemia gene from only one parent, they are considered to have a thalassemia trait. Each child of two carriers of the same type of thalassemia has a 25% chance of inheriting two copies of the mutated gene and develop the disorder; each child has a 50% chance of inheriting one mutated gene and become a carrier like their parents; and the remaining 25% chance is that each child will inherit two normal genes.

Conventional techniques for detecting thalassemia carriers, like blood and genetic testing, are costly, time-consuming, and may not be accessible to all population groups (Elshami and Alhalees, 2012). In environments with limited resources, where thalassemia is frequently prevalent, these techniques might not be practical and might not be sensitive enough to identify carriers who have subtle variations in their genetic makeup. Without the proper care and counselling, carriers may experience serious repercussions. Thus, to stop the disease from spreading to future generations, it is imperative to create novel, innovative techniques for identifying thalassemia carriers and to evaluate how well these techniques work in clinical settings.

Nevertheless, research on the effectiveness of machine learning approaches for β-thalassemia carrier detection is still lacking, and the field is still in its infancy. Therefore, it is necessary to examine the effectiveness and accuracy of machine learning approaches for β-thalassemia carrier detection, ascertain how these techniques might be incorporated into current screening programs, and evaluate the viability of applying these methods in clinical practice. Therefore, this study's results can significantly impact identifying β-thalassemia carriers and expect to develop a model as a decision-support tool, which can lead to earlier detection and intervention, improving patient outcomes and reducing the burden on healthcare systems (Subasinghe et al., 2023).

2 Methodology

The main objective of this study is to develop a time-efficient model to detect the β-Thalassemia Carriers, which can reduce the time to take a decision and develop the built model as a decision support tool.

The following subsections will discuss the step-by-step methodology of the study.

2.1 Study Setting

Data from Hemal's Adolescent and Adult Thalassemia Care Center, Mahara, Sri Lanka, one of the thalassemia treatment centres, is used in the study. All the individuals who came to the centre between August 2019 and December 2019 were considered for this study. As the population was considered for the data collection, sampling techniques have not been considered.

Table 1: Attributes of the Dataset.

Input Variable Source	Variables	Data Type
Full Blood Count	**Red Blood Cell (RBC)** count in millions per microliter (10^6/uL)	Real
	Hemoglobin (Hb) in grams per deciliter (g/dL)	Real
	Hematocrit (HCT) as a percentage (%)	Real
	Mean Corpuscular Volume (MCV) in femtolitres (fL)	Real
	Mean Corpuscular Hemoglobin (MCH) in picograms (pg)	Real
	Mean Corpuscular Hemoglobin Concentration (MCHC) in grams per deciliter (g/dL)	Real
	Red Cell Distribution Width (RDW) as a percentage (%)	Real
Hemoglobin Electrophoresis	**Hemoglobin A2 (HbA2)** as a percentage (%)	Real
Demographic Data	**Gender**	Binary
	Age	Integer
Diagnostic Report	**Phenotype** (β-Thal Carrier/β-Thal Non-Carrier)	Binary

2.2 Dataset

The original dataset consists of 343 data records. The response variable represents the phenotype, which indicates the person's carrier or non-carrier state, whereas the explanatory variables represent the haematological parameters and demographic data. The characteristics of the dataset used in this study are listed in Table 1.

2.3 Association Test

Utilising the relevant statistical tests, pre-processed data was evaluated to determine the relationships between predictor variables and the response variable at the 5% significance level and then insignificant attributes were eliminated.

2.4 Data Selection Methods and Model Fitting

This analysis is supposed to be carried out under the two data selection methods.

Method 01: Model fitting before handling the class imbalanced problem.
Method 02: Model fitting with random over-sampling technique.

Considering these two methods, the model building is carried out, and as a splitting criterion, the random split was used, then 80% of the data used as the training set, and 20% of the data used as the testing set (Das et al., 2020).

Two candidate classifiers are considered in this study: Support Vector Machine (SVM) and Probabilistic Neural Network (PNN). Using those classifiers model building was carried out under each data selection method.

2.4.1 Resampling Techniques

A classification data set with skewed class proportions is called imbalanced, leading to bias in the training dataset. This bias in the training dataset influences all classification algorithms and makes the classifier biased towards the majority class. One approach to addressing the class imbalance problem is to resample the dataset randomly. The two main approaches to randomly resampling an imbalanced dataset are to delete examples from the majority class, called under-sampling, and to duplicate examples from the minority class, called oversampling (Witten et al., 2016).

2.4.1.1 Random Oversampling

Random Oversampling includes selecting random examples from the minority class with replacement and supplementing the training data with multiple copies of those instances. Hence it is possible that a single instance may be selected multiple times (Witten et al., 2016).

2.4.2 Support Vector Machine (SVM)

A supervised machine learning approach called Support Vector Machine (SVM) is used for regression as well as classification. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane (Fig. 1). Using these support vectors, the SVM algorithm maximises the margin of the classifier. Removing the support vectors will change the position of the hyperplane (Han et al., 2012).

2.4.3 Probabilsitic Neural Network (PNN)

A probabilistic neural network (PNN) is a type of feedforward neural network that is employed to address issues related to pattern recognition and

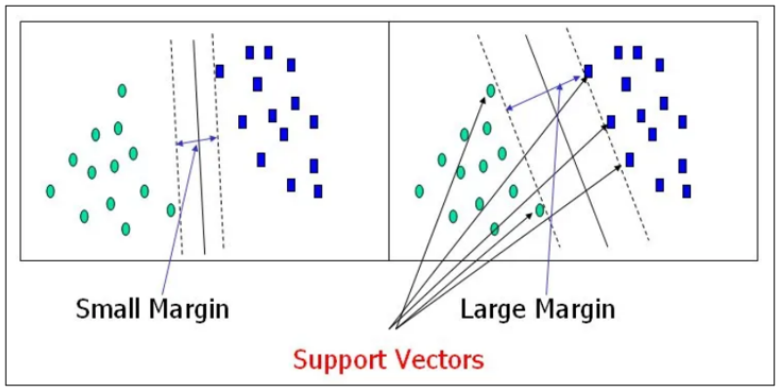

Fig. 1: Support Vectors (Source: https://images.app.goo.gl/jquBC56jebDMBmWy9).

categorisation. In the PNN technique, the parent probability distribution function (PDF) of each class is approximated using a Parzen window and a non-parametric function. The PDF of each class is then used to estimate the class probability of fresh input data, and Bayes' rule is used to allocate the class with the highest posterior probability to new input data. With this method, the possibility of misclassification is lowered. This type of artificial neural network (ANN) was created using a Bayesian network and a statistical approach known as Kernel Fisher discriminant analysis (Fig. 2) (Han et al., 2012).

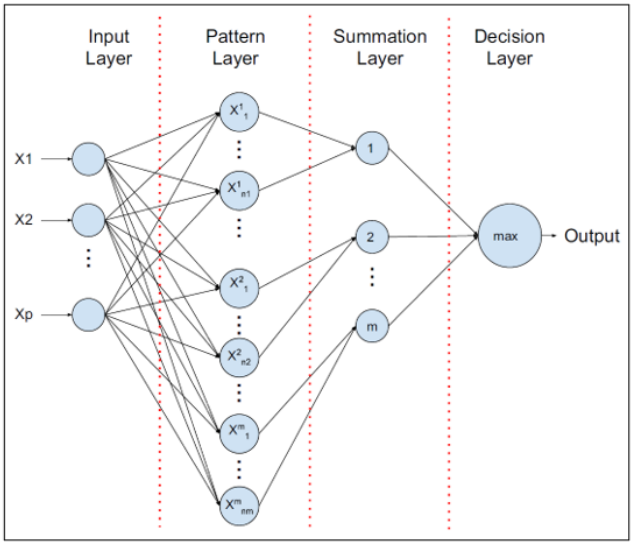

Fig. 2: Framework for Probabilistic Neural Network (Source: https://images.app.goo.gl/brpbFb1MTqtvvVWc8).

2.5 Perfomance Measures

As performance measures overall accuracy, the ability to correctly detect the β-thalassemia carriers (Sensitivity) and the ability to correctly detect the β-thalassemia non-carriers (Specificity) were considered in the analysis to identify the better models.

2.5.1 Confusion Matrix

A confusion matrix is a useful method in performance evaluation which allows to measure Accuracy, Sensitivity, and Specificity. The confusion matrix is a systematic way to allocate the predictions to the original classes to which the data originally belonged. A confusion matrix itself is also a performance evaluation technique for classification. If train a machine learning classification model on a dataset, the resulting confusion matrix will show how accurately the model categorises each record and where there might be errors (Fig. 3) (Han et al., 2012).

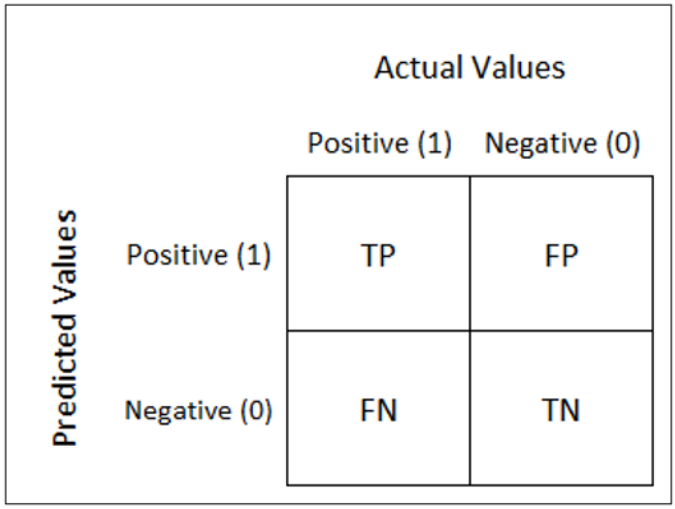

Fig. 3: Confusion Matrix.

The below diagram (Fig. 4) depicts a visual structure of the methodology.

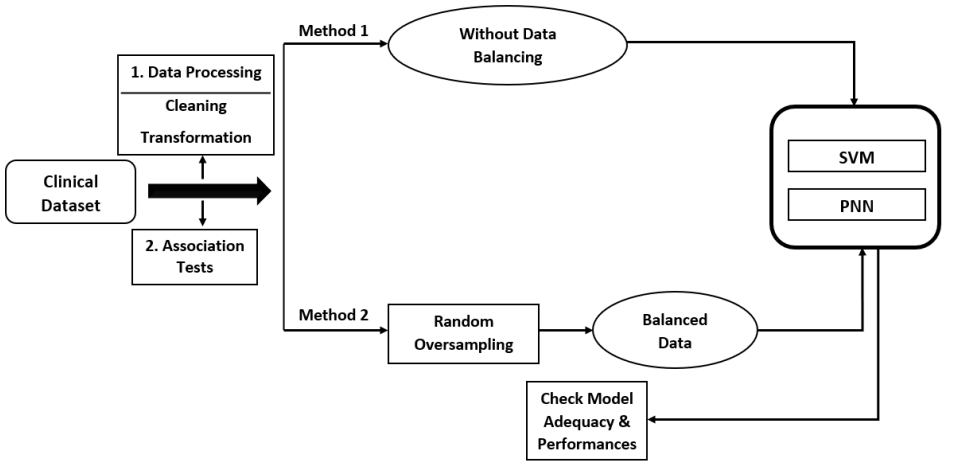

Fig. 4: Graphical Representation of the Methodology.

2.6 Ethical Statement

This study is based on the data from individuals' blood samples obtained through the centre and does not contain any personally identifiable information. Thus, it includes only secondary data. There is no significant ethical issue. Therefore, ethical clearance was not obtained for this study.

3 Results and Discussion

In the pre-processing of the dataset, missing observations were eliminated, and outlier values remained in the dataset based on the domain expert's opinion.

As illustrated in Fig. 5, following pre-processing of the original dataset, 112 (36%) people were found to be β-Thalassemia carriers, while 200 (64%) were found to be non-carriers.

Moving on to the phenotype categorization based on gender, it can be shown as follows in Fig. 6.

According to Fig. 6, it is shown that among β-Thalassemia carriers, 51% are females, and 49% are males. Also, among β-Thalassemia non-carriers, 54.5% are females, and 45.5% are males. Therefore, based on gender-wise categorization, most belong to the female category under each phenotype.

Table 2 represents the summary statistics of the used blood parameters concerning the phenotype.

Using the Chi-square and Mann-Whitney U tests, the association between the response variable and the predictor variables was assessed and indicated

Table 2: Summary Statistics of Blood Parameters.

Phenotype	Blood Parameter	Unit	Mean	Std. Dev
Beta -Thalassemia Carriers	RBC	10^6/uL	5.4	0.73
	HGB	g/dL	10.7	1.42
	HCT	%	33.6	4.25
	MCV	fL	62.1	3.66
	MCH	pg	19.7	1.19
	MCHC	g/dL	31.9	0.76
	RDW	%L	16.9	0.92
	HbA2	%	4.8	0.47
Beta -Thalassemia Non - Carriers	RBC	10^6/uL	4.72	0.48
	HGB	g/dL	12.7	1.97
	HCT	%	37.7	5.05
	MCV	fL	82.9	5.84
	MCH	pg	28.9	2.31
	MCHC	g/dL	34.2	0.77
	RDW	%L	13.3	1.51
	HbA2	%	2.5	0.25

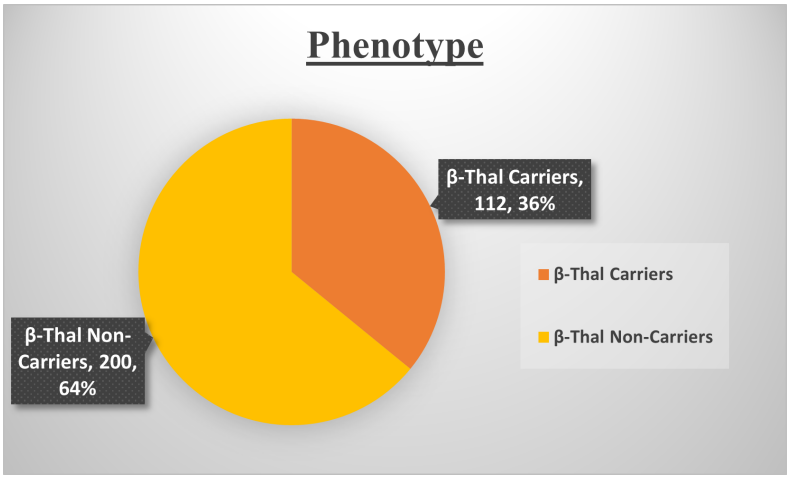

Fig. 5: Phenotype Categorization.

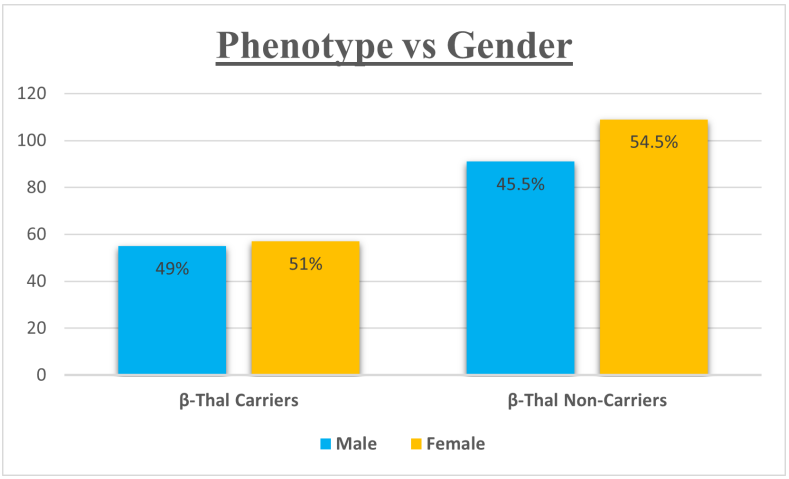

Fig. 6: Phenotype Categorization based on Gender.

that all the blood parameters are associated with the response variable except the Gender and Age variables under the 5% significance level (Table 3).

Finally, eight blood parameters: RBC, HGB, HCT, MCV, MCH, MCHC, RDW, and HbA2 were identified as associated variables with binary predictor variable (Phenotype) under the 5% level of significance.

Moving on the model fitting phase; two classification models, such as Support Vector Machine (SVM), and Probabilistic Neural Network (PNN), were designed, trained and tested, and evaluated the performances.

Machine learning based model fitting is significant since it enables the investigation and interpretation of complex relationships within the data. These methods allow for the creation of predictive models that can find hidden pat-

Table 3: Summary of Test Results for the Association Test.

Parameter	Test	P-value	Decision
RBC	Mann-Whitney	1.131e-14	Significant
HGB	Mann-Whitney	2.2e-16	Significant
HCT	Mann-Whitney	2.201e-12	Significant
MCV	Mann-Whitney	2.2e-16	Significant
MCH	Mann-Whitney	2.2e-16	Significant
MCHC	Mann-Whitney	2.2e-16	Significant
RDW	Mann-Whitney	2.2e-16	Significant
HbA2	Mann-Whitney	2.2e-16	Significant
Age	Mann-Whitney	0.1196	Not Significant
Gender	Chi-Square	0.5402	Not Significant

Table 4: Model Results of SVM Model 1 and SVM Model 2.

Method	Model	Kernel	Overall Accuracy	Ability to correctly detect the Beta-Thalassemia Carriers (Sensitivity)	Ability to correctly detect the Beta-Thalassemia Non-Carriers (Specificity)
Method 1	SVM Model 1	Linerar	89.32%	90.3%	85.74%
Method 2	SVM Model 2	Linerar	94.5%	95.6%	95%

terns, create precise forecasts, and extract insightful information (Padhy et al., 2012).

3.1 Support Vector Machine (SVM) Models

Three different kernels were used to assess performance during the Support Vector Machine (SVM) model fitting process: Linear, Polynomial, and Radial Basis Function (RBF). The SVM model's performance and suitability for the specific task were assessed in-depth by fitting the model with several kernels. According to that, the best performances were given by the models fitted under the Linear kernel (Table 4).

When considering the SVM Model 1, its overall accuracy is 89.32%. Under Method 2, fitted SVM Model 2 shows 94.5% overall accuracy and indicates higher sensitivity than SVM Model 1. Compared to these two models, better results are given by the model fitted under Method 2, SVM Model 2.

Table 5: Model Results of PNN Model 1 and PNN Model 2.

Method	Model	Spread	Overall Accuracy	Ability to correctly detect the Beta-Thalassemia Carriers (Sensitivity)	Ability to correctly detect the Beta-Thalassemia Non-Carriers (Specificity)
Method 1	PNN Model 1	0.5	91.9%	94.4%	90.9%
Method 2	PNN Model 2	0.6	98.75%	96.42%	92.9%

3.2 Probabilistic Neural Network (PNN) Models

The next approach was to fit a Probabilistic Neural Network (PNN) model. To optimize the performance of the PNN model, different spread (hyper-parameter) values were used when fitting the model.

Based on that, Table 5 indicates that the PNN Model 1 fitted with a spread of 0.5, which has an overall accuracy of 91.9%, had the best result under Method 1. Likewise, the PNN Model 2 fitted using Method 2 had superior outcomes, with a spread of 0.6, reflecting a 98.75% overall accuracy and 96.42% sensitivity. When these two PNN models from the two approaches were compared, the PNN Model 2 that was fitted using Method 2 performed better.

The following Fig. 7 shows the network architecture of PNN Model 2 obtained under the model fitting.

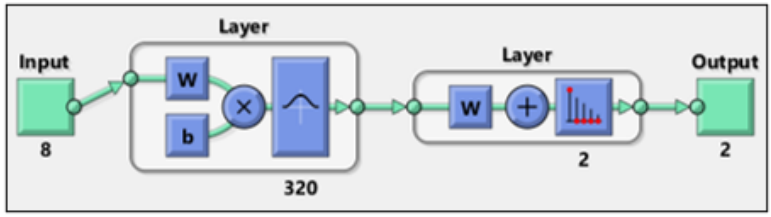

Fig. 7: Network Architecture of PNN Model 2

The input layer has eight nodes, which implies that eight predictor variables are added. 320 records in the training set are matched to a pattern layer with 320 nodes. The summing layer then represents the two nodes that make up the target variable's two categories. Subsequently, the output layer pairs each target category's weighted vote from the pattern layer, using the vote

Table 6: Results of SVM and PNN Models' Performances.

Model	Overall Accuracy	Ability to correctly detect the Beta-Thalassemia Carriers (Sensitivity)	Ability to correctly detect the Beta-Thalassemia Non-Carriers (Specificity)
SVM Model 1	89.32%	90.3%	85.74%
SVM Model 2	94.5%	95.6%	95%
PNN Model 1	91.9%	94.4%	90.9%
PNN Model 2	98.75%	96.42%	92.9%

with the highest percentage to forecast the target category. Afterwards, the two nodes are represented by the output layer.

All of the models fitted using Method 2 (model fitting with random over-sampling technique) yielded better results when the performance of the fitted models was evaluated (Table 6). As a result, the random over-sampling approach used in Method 2 significantly enhanced model functionality, reduced class imbalance problems, and ultimately led to better study findings.

3.3 Selecting the Better Model

To find a better model to detect the β-thalassemia carriers, the performances of all the fitted models were compared and suggest a better-performing model. Table 6 represents all the fitted models' results obtained in this study.

Every model performs better, as seen in Table 6 models' performances. SVM Model 2 and PNN Model 2 show even greater performance among them. In terms of accuracy in identifying carriers of β-thalassemia, PNN Model 2 outperforms SVM Model 2 by a small margin. Also, PNN Model 2 shows better values for sensitivity (96.42%) and overall accuracy (98.75%). Taking into account these results, PNN Model 2 fitted using Method 2 can be recommended as a more accurate model for identifying β-thalassemia carriers.

Some limitations of the study are identified as follows: even though there are several types of thalassemia, only β-Thalassemia carriers were considered, the blood test measures the different parameters in the blood, but the most significant parameters were used for this study based on the domain expert's knowledge.

As future work, advanced resampling methods can be employed to overcome the class imbalanced problem.

4 Conclusion

This study aims to develop a time-efficient model to detect the β-Thalassemia Carriers in Sri Lanka, which can reduce the time to take a decision and develop the built model as a decision support tool. The following conclusions were drawn after fitting the appropriate models to achieve the aforementioned objective.

All blood parameters, including RBC, HGB, HCT, MCV, MCH, MCHC, RDW, and HbA2, were associated with the Phenotype variable concerning the considered dataset at a 5% level of significance. Furthermore, as compared to the models fitted using Method 1 (model fitting prior to addressing the class imbalance issue), all of the models built using Method 2, i.e. model fitting with random over-sampling, produced better results. At the same time, the selected machine learning models; Support Vector Machine (SVM) and Probabilistic Neural Network (PNN) models performed well in classifying the β-Thalassem-ia carrier or non-carrier state, and the PNN model performed even better than the SVM model. Finally, considering the overall and classification accuracies, the PNN model fitted with random over-sampling (PNN Model 2) performed better for detecting β-thalassemia carriers. Furthermore, that model can be developed into a tool for decision-support, assisting doctors in their decision-making.

Acknowledgement
The authors extremely would like to acknowledge the support extended by Hemal's Adolescent and Adult Thalassemia Care Center, Mahara, Sri Lanka for providing the dataset.

References

CDC (2023). Thalassemia, URL: https://www.cdc.gov/thalassemia/about/index.html

Das R. et al. (2020). A decision support scheme for beta-thalassemia and HbE carrier screening. *Journal of Advanced Research*. 24: 183-190. DOI: 10.1016/j.jare.2020.04.005

Elshami E. H. and Alhalees A. M. (2012). Automated Diagnosis of Thalassemia Based on DataMining Classifiers. *The International Conference on Informatics & ApplicationsAt: University Sultan Zainal Abidin, Malaysia (2012)*. DOI: 10.13140/RG.2.1.3336.1362

Han J., Kamber M., and Pei J. (2012). *Data Mining Concepts and Techniques*. (3rd ed.). Elsevier Inc.

Mohammed M. Q. and Al-Tuwaijari J. M. (2021). A Survey on various Machine Learning Approaches for thalassemia detection and classification.

Turkish Journal of Computer and Mathematics Education. 12(13): 7866-7871. DOI:10.17762/turcomat.v12i13.11284

NHS - Thalassemia (2022). URL: https://www.nhs.uk/conditions/thalassaemia/#:̃:text=A%20carrier%20of%20thalassaemia%20is,you%20will%20not%20develop%20thalassaemia.

Padhy N., Mishra P. and Panigrahi R. (2012). The Survey of Data Mining Applications And Feature Scope abs/1211.5723. *ArXiv*. URL: https://api.semanticscholar.org/CorpusID:2992477

Patel P. et al. (2019). Examining depression and quality of life in patients with thalassemia in Sri Lanka. *International Journal of Noncommunicable Diseases*. 4(1): 27-33. DOI: 10.4103/jncd.jncd_49_18

Saritas M. and YASAR A. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *International Journal of Intelligent Systems and Applications*. 7(2): 88–91. DOI: 10.18201//ijisae.2019252786.

Subasinghe G.K., Chandrasekara N.V. and Premawardhena A.P. (2023). Exploring data mining avenues in β-Thalassemia carrier identification. *Proceedings of the International Conference on Applied and Pure Sciences (ICAPS 2023-Kelaniya) Volume 3, Faculty of Science, University of Kelaniya Sri Lanka. Page 95*. URL: http://repository.kln.ac.lk/handle/123456789/26930

Witten I. H., Frank E. and Hall M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. (3rd ed.). Elsevier Inc.