AUIQ Technical Engineering Science

Manuscript 1021

Investigating the Differential Effects of SMOTE Variants on Class Imbalance and Exploring Their Applicability to a Thalassemia Prediction Model

Hussam Mezher Merdas

Ayad Hameed Mousa

AUIQ TECHNICAL ENGINEERING SCIENCE 2025;2:9–19

ORIGINAL STUDY

Investigating the Differential Effects of SMOTE Variants on Class Imbalance and Exploring Their Applicability to a Thalassemia Prediction Model

Hussam Mezher Merdas [a,*], **Ayad Hameed Mousa** [b]

[a] Department of Accounting, College of Management and Economics, University of Warith Al-Anbiyaa, Kerbala, Iraq
[b] Department of Computer Science, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq

ABSTRACT

Researchers work around the clock on many datasets provided by various institutions. These researchers strive to come up with highly efficient Artificial Intelligence models. Often, researchers face the problem of imbalance in the distribution of classes in a particular feature in the selected dataset, which creates an Artificial Intelligence model biased towards one class at the expense of another class that is no less important than the first. On the other hand, thalassemia is a disease that affects people of different ages. The degree of disease varies according to the thalassemia class. This study proposes an improved Machine Learning model that aims to provide a comprehensive comparison of different SMOTE techniques to enhance class balance in thalassemia prediction. And create a Machine Learning model that predicts the possibility of an individual suffering from thalassemia based on the data modified by the proposed SMOTE technology. This study concluded, according to the proposed model, that the best SMOTE technique that can be used in such datasets with clear imbalance is the SMOTE-ENN technique, as the model achieved high-accuracy prediction results, as the model's accuracy was 99% and the F1-score was 97%. This study provides software developers with the steps and source code to develop it as a mobile or computer application to help people know the probability of their infection with thalassemia. The study also helps researchers determine the best SMOTE technique that is compatible with imbalanced datasets.

Keywords: Artificial intelligence, Machine learning, SMOTE, Thalassemia

1. Introduction

The dataset is the basic building block of every artificial intelligence model that relies on previous training data. Classification is also one of the most important branches of machine learning [1]. To perform the classification process, datasets are used that usually suffer from the problem of imbalance. The reason for this condition is that some categories in the dataset are significantly lower than other categories, which generates an imbalanced dataset. An example of this is a dataset for a specific disease that contains two categories, the first of which is related to the possibility of a person getting the disease and the other of the possibility of not getting it. It is noted that the second category is much lower than the first, which causes the imbalance problem. To address the data imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is used. This technique has several types that will be discussed in detail later in this article. In this article, the focus was on thalassemia, which is one of the dangerous diseases that require study. The classification technique was relied upon to predict the probability of a person getting this disease later based on past data and factors. Thalassemia occurs as a result of a genetic disorder in the blood cells. It can be defined as a decrease in the level of hemoglobin in the blood and

Received 13 December 2024; revised 10 March 2025; accepted 22 March 2025.
Available online 30 April 2025

* Corresponding author
E-mail addresses: hussam.mezher@uowa.edu.iq (H. M. Merdas), ayad.h@uokerbala.edu.iq (A. H. Mousa).

https://doi.org/10.70645/3078-3437.1021

a decrease in the number of red blood cells. The most important symptoms of the disease are pallor, fatigue, and exhaustion resulting from the decrease in the level of hemoglobin in the blood, which is responsible for transporting oxygen in the body. This disease is a genetic disease resulting from changes in the cells that form hemoglobin [2]. This genetic mutation causes a deficiency in the production of hemoglobin in the blood and a decrease in the number of red blood cells, causing symptoms of anemia [3]. To prevent this disease, it is recommended to conduct the necessary tests before marriage to determine the possibility of contracting this disease [4]. Modern artificial intelligence techniques help to know in advance the possibility of a person contracting this disease through datasets provided by health centers concerned with this disease. This study focuses on the possibility of predicting the occurrence of thalassemia based on machine learning algorithms and based on previous data. This study relied on the classification technique, which is one of the supervised learning techniques, as the model is trained on only two categories, which are the occurrence or non-occurrence of the disease. Several studies have addressed the prediction of the occurrence of thalassemia as well as the use of classification and the use of SMOTE technology, including Ananina Devanath and others have proposed a model based on machine learning algorithms, and SMOTE technology has been employed to balance the data used. They used the following algorithms: K-Nearest Neighbor (kNN), Logistic Regression, Adaptive Boosting (ADA Boosting), Multilayer Perceptron (MLP), Gradient Boosting classifier, and others. The model gave 100% prediction accuracy using the ADA Boosting algorithm [5]. Alaa S. AlAgha and others designed a model that predicts the occurrence of β-thalassemia. The model had two stages, the first in which the data was balanced using SMOTE, and in the second stage several artificial intelligence algorithms were used, including k-nearest neighbors (k-NN), naïve Bayesian (NB), decision tree (DT) and the multilayer perceptron (MLP) neural network. The model gave a prediction accuracy of approximately 99% [6]. Muniba Saleem and his group proposed a model based on two pillars. The first is employing five feature selection approaches to select the best features that affect prediction. The second pillar was the employment of nine algorithms. Also, to obtain good prediction accuracy, SMOTE was used. The results were good with an accuracy of 93% [7].

The three research studies above employed one type of SMOTE with several artificial intelligence algorithms.

This study, for the first time, employs several types of SMOTE and uses a single machine-learning algorithm. The steps of the model are as follows: feeding the model with a local Iraqi dataset for thalassemia. We work to balance the classes using several types of SMOTE, enter the result of each type into the Random Forest algorithm, and then measure the prediction accuracy in each case to find out which type of SMOTE is the best. Although previous studies have applied SMOTE for handling class imbalance in medical datasets, most have used a single variant without systematically comparing different approaches. This study addresses this gap by evaluating multiple SMOTE variants and identifying the most effective technique for improving thalassemia classification accuracy. This comparison is crucial as different SMOTE techniques may have varying impacts depending on dataset characteristics.

2. Materials and methods

The proposed model in this study was designed to be used and developed as an application that can be installed on a mobile or computer. The proposed model can be integrated into healthcare settings in several ways:

For Doctors: It can assist physicians in identifying at-risk patients by providing automated predictions based on routine blood tests .For Patients: A user-friendly mobile application could allow individuals to enter their blood test results and receive an initial assessment of their risk for thalassemia. For Researchers: The model can be used as a tool for analyzing large-scale medical datasets to study disease prevalence and improve classification techniques. This model consists of several steps. The first step is to enter the dataset for thalassemia disease, where the necessary pre-processing was done to produce an organized and suitable dataset for the work. Then the data set was entered into the Random Forest algorithm, and then the results obtained from the algorithm were displayed. After that, the process was repeated, but this time the SMOTE technique was used to balance the dataset, and then it was entered again into the Random Forest algorithm, and the results were displayed again. Also, for the third time, the dataset was entered on several different types of SMOTE techniques to balance the dataset and for each type of SMOTE separately, and the dataset was entered with each type into the Random Forest algorithm. In the last stage, the results obtained from the three stages were compared. Finally, the best technique that gives the most accurate prediction result

was determined. For more clarification, Fig. 1 can be seen below.

2.1. Dataset and preprocessing

The dataset used in this study was obtained from Al-Hussein Medical City in Karbala Governorate in Iraq. This dataset consists of (9) columns and (1380) rows for thalassemia patients who visited the hospital. These rows were as follows: (ID, gender, RBC, HGB, HCT, MCV, MCH, MCHC, RDW) where RBC stands for Red Blood Cell and the normal level is between 3.50 and 5.50 units (10^12/L). While HGB is an abbreviation for Hemoglobin and the normal level of it in the blood is 11.0 to 16.0, Unit: (g/dL). As for HCT, it is an abbreviation for Hematocrit which represents the proportion, by volume, of the Blood that consists of red blood cells, and the normal level is between 36.0 to 48.0 percentage units. As for MCV, it is the Mean Corpuscular Volume and the normal level is between 80.0 to 99.0, Unit: fL. MCH stands for Mean Corpuscular Hemoglobin, which is the average amount of hemoglobin in the average red cell. The normal range is between 26.0 and 32.0, Unit: pg. MCHC stands for Mean Corpuscular Hemoglobin Concentration, which is the normal range between 32.0 and 36.0 units (g/dL). RDW stands for Red Blood Cell Distribution Width, which is the normal range between 11.5 and 14.5 units, measured in percentages [8, 9].

The dataset that was initially obtained lacked the most important column, which is (target). To generate this column, a programming function was created within the model proposed in this study. The software function divided the dataset into five main categories based on accurate medical information, which are as follows: (Beta-Thalassemia Minor, Beta-Thalassemia Major, Alpha-Thalassemia Minor, Alpha-Thalassemia Major, Normal). When the MCV is between (60–70) and MCH is between (19–23) and HGB is between (6–11) and RBC is between (4.5–6.3), this means that the patient suffers from Minor Beta-Thalassemia. When the MCV is between (50–70) and MCH is between (12–20) and HGB is between (6–11) and RBC is between (4.5–6.3), this means that the patient suffers from Major Beta-Thalassemia. When the MCV is less than or equal to 79 and MCH is less than 27 and HGB is between (7–10) and RBC is between (4.5–6.3), this means that the patient suffers from Minor Alpha-Thalassemia. When MCV is less than or equal to 79, MCH is less than 27, HGB is between (5–6), and RBC is between (4.5–6.3), this means that the patient suffers from Major Alpha-Thalassemia. Otherwise, the function classifies the data as normal.

For all of the above, the target attribute now consists of four main categories that include two different types of thalassemia, each with a specific level (minor or major). In addition to these four categories, there is a fifth category, which is normal, meaning that the patient's condition is healthy. Fig. 2 below shows the levels of these categories. Pre-processing operations were performed, which consisted of examining the dataset to see if it contained empty fields, and the necessary treatments were taken to obtain a complete dataset. After that, the dataset was examined to see if there was duplicate data, and after examination, it was found that there was no duplicate. After performing all of the above operations, the dataset became ready to be entered into the Random Forest algorithm, which will be explained in the next stage of this study.

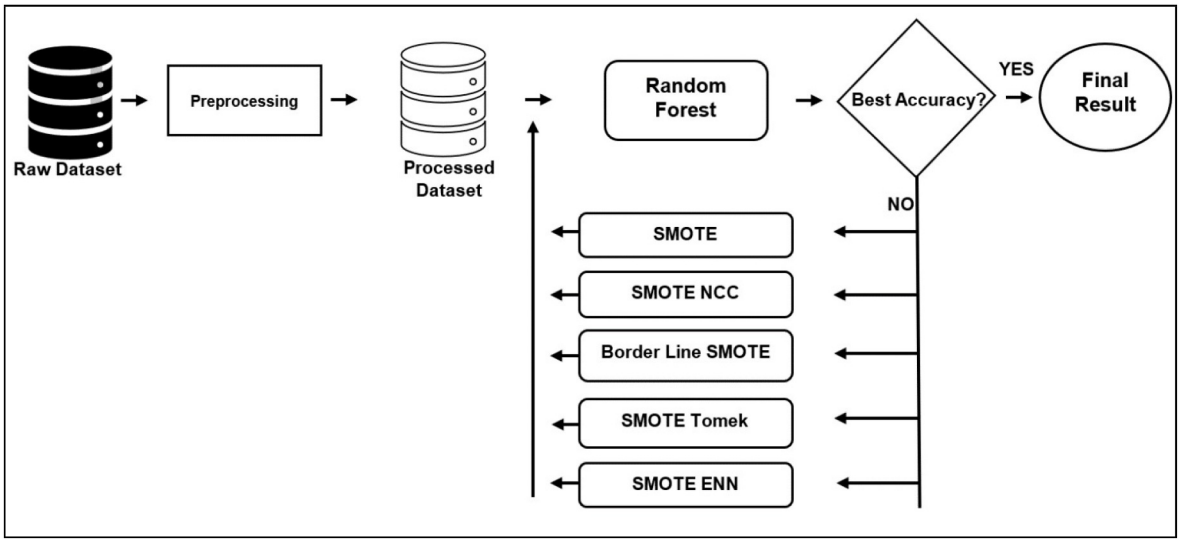

Fig. 1. General steps.

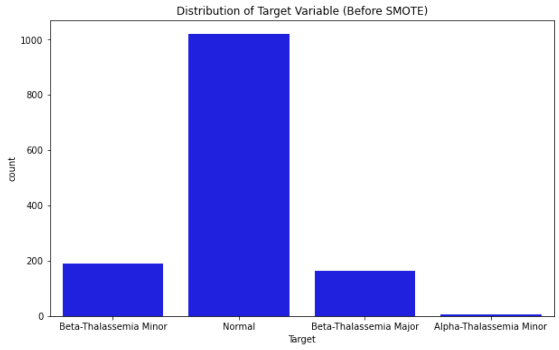

Fig. 2. Distribution of target variables in dataset.

2.2. Random forest algorithm

Random Forest is a machine learning algorithm used in classification or regression [1]. It works by creating multiple decision trees during training and outputting the class pattern (classification) or average prediction (regression) for the individual trees. This algorithm essentially consists of multiple decision trees [10]. Decision trees are streamlined with internal branches that represent the test and terminal leaves that represent the results. Random Forest is one of the most important ensemble learning techniques, which means that it builds multiple models (decision trees) and combines their results to improve performance. In classification tasks, the algorithm chooses the class that gets the most votes from among all the trees [11].

This algorithm also relies on the use of random samples, as it uses Bagging (Bootstrap Aggregating): where it takes random samples from the dataset with replacement to train each decision tree. This leads to introducing diversity among the trees, which reduces overfitting.

The random forest algorithm relies on a method represented by randomization of features: while dividing the nodes, the random forest does not evaluate all possible features but selects a random subset of features to divide on. This randomness further decorates the trees and improves performance [1].

The classification steps are this algorithm as follows: The first stage is data preparation: Divide the data into training and test sets. Then comes the tree-building stage where multiple decision trees are created, each of which is trained on a preliminary sample of the data (random subsets of the training data). At each node in the tree, a random subset of features is selected, and the best feature from this subset is used to perform the division [12].

Then it comes to the voting stage: After creating the tree forest, a new sample is classified by passing it through all the trees.

Then each tree gives a classification, and the final prediction is made based on majority voting (i.e. the class that is most often predicted by the trees is chosen) [12].

In summary, the random forest is a powerful classification algorithm due to its ability to handle large datasets, reduce overfitting, and provide high accuracy. This algorithm is particularly effective when you need a reliable and flexible model that performs well even with noisy or missing data.

In the proposed model, the Random Forest algorithm was the core of the model. Random Forest was chosen due to its robustness in handling imbalanced datasets, ability to capture complex feature interactions, and resilience against overfitting. Compared to SVM and XGBoost, Random Forest requires less hyperparameter tuning and performs well with small to moderately sized datasets, making it an ideal choice for this study. Additionally, prior research on medical classification problems has demonstrated its effectiveness in producing high accuracy with imbalanced data. However, the feature that the algorithm will use for prediction, which is targit, was unbalanced in the number of values. Therefore, SMOTE techniques will be used later to create the required balance. However, at this stage, the dataset was entered directly into this algorithm without balancing the data in order to observe the results obtained and compare them when using several types of SMOTE in the later stages. Fig. 3 below shows the obtained results.

Precision measures the proportion of positive identifications that were actually correct. High precision indicates that the classifier has a low rate of false positives. It's particularly important in scenarios where the cost of false positives is high. Its mathematical equation is [13]:

Classification Report :-

	precision	recall	f1-score	support
Normal	1.00	0.99	1.00	198
Beta-Thalassemia Minor	1.00	1.00	1.00	30
Beta-Thalassemia Major	0.96	1.00	0.98	47
Alpha-Thalassemia Minor	0.00	0.00	0.00	1
accuracy			0.99	276
macro avg	0.74	0.75	0.74	276
weighted avg	0.99	0.99	0.99	276

Confusion Matrix :-

```
[[197   0   1   0]
 [  0  30   0   0]
 [  0   0  47   0]
 [  0   0   1   0]]
```

Fig. 3. Results obtained from the random forest algorithm before applying the SMOTE technique.

$$\text{Precision} = \frac{\textit{True Positive } (TP)}{\textit{True Positive } (TP) + \textit{False Positive } (FP)} \quad (1)$$

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives that were correctly identified. It answers the question: Of all actual instances of a particular class, how many were correctly predicted? High recall indicates that the classifier successfully captures most of the positive instances. It's crucial in situations where missing a positive instance is costly. Its mathematical equation is [13]:

$$\text{Recall} = \frac{\textit{True Positive } (TP)}{\textit{True Positive } (TP) + \textit{False Nagative } (FN)} \quad (2)$$

The F1-score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall. The F1-score is useful when you need a balance between precision and recall, especially when you have uneven class distributions. Its mathematical equation is [13]:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Support refers to the number of actual occurrences of each class in the dataset. It indicates how many samples belong to each class. Support is useful for understanding the distribution of classes and the context of precision, recall, and F1-score. It helps in assessing whether the metrics are influenced by class imbalance.

The macro average computes the metric independently for each class and then takes the average, treating all classes equally regardless of their support. Macro averaging is useful when you want to evaluate the model's performance equally across all classes, without considering class imbalance. It can be heavily influenced by classes with fewer instances. Its mathematical equation is [14]:

$$\text{Macro Precision} = \frac{\sum \text{Precision}_i}{N} \quad (4)$$

Where N is the number of classes.

The weighted average computes the metric for each class and then takes the average, weighted by the number of instances (support) of each class. Weighted averaging accounts for class imbalance by giving more weight to classes with more instances. It's useful when you want an overall performance metric that reflects the distribution of the dataset. Its mathematical equation is [14]:

$$\text{Weighted Precision} = \frac{\sum (\text{Precision}_i \times \text{Support}_i)}{\text{Support}_i}) \quad (5)$$

By observing Fig. 3 above, it is clear that the prediction accuracy of the Random Forest algorithm was 99%. By observing the confusion matrix, it is clear that the model works well with the categories Normal, Beta-Thalassemia Minor, and Beta-Thalassemia Major, as it predicted most of the cases correctly, but it is clear that the model completely failed to classify the last category, Alpha-Thalassemia Minor, as it took one case and classified it as Beta-Thalassemia Minor, contrary to reality. The main reason for this is the imbalance of data in the (Target) column, which is illustrated in Fig. 4 below, which shows that the Alpha-Thalassemia Minor category had the least data size, as it contained only 7 cases. On this basis, this model was proposed, in which the most important types of SMOTE will be used to achieve the required balance and thus obtain an ideal prediction for thalassemia disease, which will be explained in the following steps in this study.

2.3. SMOTE

SMOTE (Synthetic Minority Oversampling Technique) is a powerful technique used to balance data within a single feature in a dataset. Imbalanced data occurs when the number of instances in one class is significantly higher than the rest of the classes, which can negatively impact the performance of machine learning models [15]. This technique helps balance the dataset by creating synthetic samples for the minority class. SMOTE works by identifying the class with the least samples, known as the minority class. SMOTE selects one or more nearest neighbors of each sample from the minority class (based on distance measures such as Euclidean distance). It then creates new synthetic data points by interpolating between the original sample and its selected neighbor [16]. This means that the new data point will lie somewhere along the line between the original sample and

```
Normal                    1019
Beta-Thalassemia Minor     188
Beta-Thalassemia Major     165
Alpha-Thalassemia Minor      7
Name: Target, dtype: int64
```

Fig. 4. Size of target feature classes in the dataset.

its neighbor in the feature space. These newly generated synthetic samples are then added to the dataset to balance the ratio between the minority and majority classes. The SMOTE technique was employed in the proposed model and indeed it generated balanced data and as Fig. 5 below shows how the classes are distributed.

The data generated by SMOTE were entered into the Random Forest algorithm and gave the results indicated in Fig. 6 below. It is clear from this figure that the algorithm gave a prediction accuracy of 98%, which is less than the previous result, but it is noticeable that the model predicted the Alpha-Thalassemia Minor case well, as out of 208 cases, it failed to predict only 16 cases, giving a percentage of 96% according to the f1-score measure, and this is noticeable progress in the model's performance if compared to the previous results in which the f1-score was zero. To obtain better results, other types of SMOTE were used, which will be explained in detail with their results in this study.

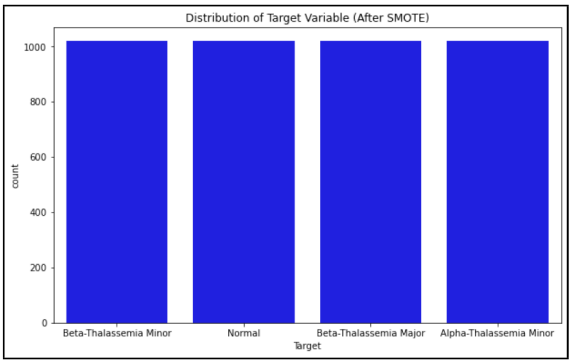

Fig. 5. Distribution of target variables in dataset (after SMOTE).

Classification Report :-

	precision	recall	f1-score	support
Normal	1.00	1.00	1.00	223
Beta-Thalassemia Minor	0.97	0.99	0.98	196
Beta-Thalassemia Major	0.95	0.99	0.97	173
Alpha-Thalassemia Minor	0.99	0.93	0.96	224
accuracy			0.98	816
macro avg	0.98	0.98	0.98	816
weighted avg	0.98	0.98	0.98	816

Confusion Matrix :-

```
[[222   0   0   1]
 [  0 195   0   1]
 [  0   0 172   1]
 [  0   6  10 208]]
```

Fig. 6. Results obtained from the random forest algorithm after applying the SMOTE technique.

2.4. SMOTENC

SMOTENC (Synthetic Minority Oversampling Technique for Nominal and Continuous Attributes) is a variant of the SMOTE technique specifically designed to handle datasets that contain both categorical (nominal) and numerical (continuous) attributes [17]. In many real-world scenarios, datasets contain a mix of these attribute types, and the standard SMOTE method, which only works on numerical data, cannot be directly applied. SMOTENC addresses this problem by treating the two attribute types differently. The technique can be summarized in several lines: For numeric attributes, SMOTENC works in the same way as SMOTE, using nearest neighbors and interpolation to create synthetic samples among the minority class instances. For categorical attributes, categorical attributes cannot be interpolated like numerical attributes. Instead, SMOTENC uses the mode (most frequent value) of the categorical features from the nearest neighbors of a given sample. When creating a synthetic instance, the value that appears most frequently among the neighbors is assigned to the categorical feature instead of numerical interpolation. SMOTENC can also process both categorical and continuous features at the same time, allowing it to create synthetic samples that respect the nature of both types of data [18]. The SMOTE technique was employed in the proposed model and indeed it generated balanced data and as Fig. 7 below shows how the classes are distributed.

When the data generated by SMOTENC were fed into the Random Forest algorithm, the results were close to the previous results using the standard SMOTE technique. Except for the prediction of the second category, Beta-Thalassemia Minor, the current results were slightly better, giving a percentage of 99% according to the f1-score measure compared to 98% when using the standard SMOTE technique. Fig. 8 below illustrates this.

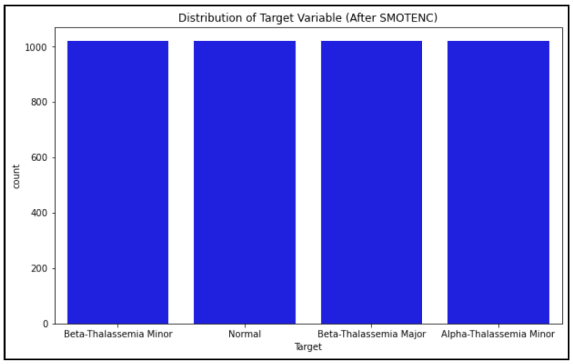

Fig. 7. Distribution of target variables in dataset (after SMOTENC).

```
Classification Report :-
                    precision    recall  f1-score   support

Normal                   1.00      1.00      1.00       223
Beta-Thalassemia Minor   1.00      0.97      0.99       196
Beta-Thalassemia Major   0.94      1.00      0.97       173
Alpha-Thalassemia Minor  0.98      0.95      0.96       224

    accuracy                           0.98       816
   macro avg       0.98      0.98      0.98       816
weighted avg       0.98      0.98      0.98       816


Confusion Matrix :-
[[223   0   0   0]
 [  0 191   0   5]
 [  0   0 173   0]
 [  0   0  11 213]]
```

Fig. 8. Results obtained from the random forest algorithm after applying the SMOTENC technique.

2.5. Borderline-SMOTE

Borderline-SMOTE (Synthetic Minority Oversampling Technique) is a variant of SMOTE that addresses the problem of class sizes differing within a single feature in a dataset. This technique focuses on "borderline" cases, which are the minority class cases closest to the decision boundary separating the two classes. These cases are more likely to be misclassified because they are close to the majority class. This creates a weak model that is biased toward the majority class because it has difficulty correctly classifying these cases, resulting in poor performance. Borderline-SMOTE works by identifying borderline cases [19]. First, it identifies minority class cases close to the decision boundary or majority class cases based on the number of nearest neighbors that belong to the majority class. The second step is to oversample near the boundary. Instead of creating synthetic samples throughout the minority class space, Borderline-SMOTE focuses on these borderline cases. The assumption is that synthetic data near the decision boundary will help the model better distinguish between classes. Meanwhile, there are two versions Borderline-SMOTE1: which generates synthetic samples of the marginal minority cases, and Borderline-SMOTE2: which generates synthetic samples of the marginal minority cases, in addition to introducing some majority class cases to reduce overfitting. The advantage of Borderline-SMOTE is that it focuses on the "most important" minority class cases (those close to the decision boundary), which are more likely to be misclassified. It also improves the performance of the classifier on imbalanced datasets without overwhelming the classifier with unnecessary synthetic data [20]. By focusing on the marginal cases, this method tends to produce better results than the original SMOTE method in cases where misclassification of minority class cases close to the decision boundary is a major concern. The distribution of the Target feature classes in the dataset generated by this technique is shown in Fig. 9 below.

Fig. 10 below shows the results obtained from this technique after entering the resulting data into the random forest algorithm. It is noticeable that the results are not much different from the two previous techniques, except that this technique gave greater accuracy than its predecessors in predicting the rare category represented by Alpha-Thalassemia Minor, as the accuracy according to the f1-score measure was 97%. As it failed to classify only 6 cases out of 218 cases.

2.6. SMOTE-Tomek

SMOTE-Tomek This technique differs from its predecessors in that it is a hybrid technique that combines SMOTE and Tomek links [21]. The SMOTE technique was previously explained in this study,

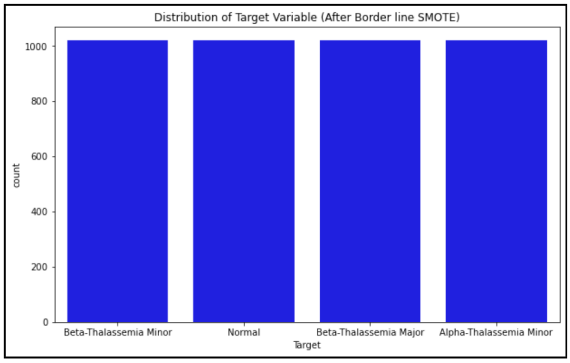

Fig. 9. Distribution of target variables in dataset (after Borderline-SMOTE).

```
Classification Report :-
                    precision    recall  f1-score   support

Normal                   1.00      1.00      1.00       223
Beta-Thalassemia Minor   0.98      0.98      0.98       196
Beta-Thalassemia Major   0.98      0.98      0.98       173
Alpha-Thalassemia Minor  0.97      0.97      0.97       224

    accuracy                           0.98       816
   macro avg       0.98      0.98      0.98       816
weighted avg       0.98      0.98      0.98       816


Confusion Matrix :-
[[222   0   0   1]
 [  0 193   0   3]
 [  0   0 170   3]
 [  0   3   3 218]]
```

Fig. 10. Results obtained from the random forest algorithm after applying the borderline-SMOTE technique.

while Tomek links are a data-cleaning technique used to remove noisy samples from a dataset, especially those close to the decision boundaries between the majority and minority classes. A Tomek link is a pair of instances (one from the minority class and one from the majority class) that are the closest neighbors to each other and are from different classes. The idea is that such pairs are likely to be borderline or noisy cases, and removing the majority class instance from these pairs can help clarify the decision boundaries [21]. SMOTE-Tomek works by oversampling using SMOTE, where SMOTE is applied to generate synthetic samples for the minority class, increasing their size to balance the dataset. Then comes the cleaning stage using Tomek links, where after oversampling, Tomek links are identified. For each Tomek link, the majority of the class instance in the pair is removed, which helps clean the dataset and reduces potential noise or overlap between classes. One of the most important features of this technique is creating a balanced dataset. SMOTE helps balance the dataset by oversampling the minority class, making it easier for the model to learn patterns for both classes. By applying Tomek links after SMOTE, the method removes noisy or marginal majority class samples that might confuse the model. This results in clearer decision boundaries between classes, which improves classification performance. Overall, SMOTE-Tomek enhances the ability of machine learning models to handle imbalanced datasets by combining the strengths of SMOTE (oversampling) and Tomek links (data cleaning) [22]. Fig. 11 below shows the distribution of classes resulting from this technique.

The results obtained from this technique, as shown in Fig. 12 below, did not differ much from its predecessors, as they are good results with a prediction accuracy of 98%. To obtain excellent accuracy and good distribution, other techniques will be taken, as the last technique in this study will meet this purpose.

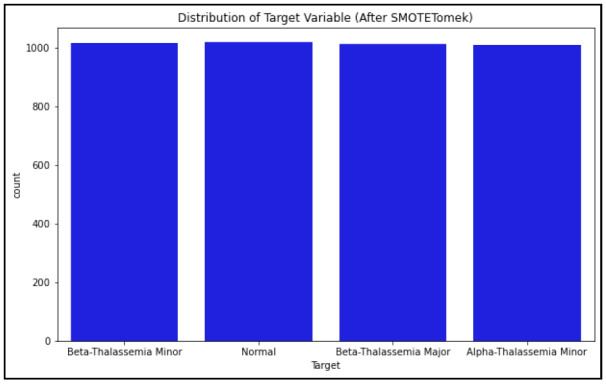

Fig. 11. Distribution of target variables in dataset (after SMOTE-Tomek).

```
Classification Report :-
                        precision    recall  f1-score   support

Normal                       1.00      0.99      1.00       203
Beta-Thalassemia Minor       0.99      0.99      0.99       205
Beta-Thalassemia Major       0.96      0.99      0.98       199
Alpha-Thalassemia Minor      0.97      0.95      0.96       204

    accuracy                                     0.98       811
   macro avg                 0.98      0.98      0.98       811
weighted avg                 0.98      0.98      0.98       811


Confusion Matrix :-
[[201   0   0   2]
 [  0 202   0   3]
 [  0   0 197   2]
 [  0   2   8 194]]
```

Fig. 12. Results obtained from the random forest algorithm after applying the SMOTE-Tomek technique.

2.7. SMOTE-ENN

SMOTE-ENN is a hybrid technique like its predecessor but combines SMOTE and ENN (Edited Nearest Neighbors) to address the problem of imbalanced datasets [23]. It improves the performance of the model by oversampling the minority class and removing noisy or misclassified instances from the dataset. SMOTE-ENN works by employing SMOTE to generate synthetic samples of the minority class by identifying a minority class instance and interfering between it and its nearest neighbors within the same class. This oversampling method helps balance the dataset by increasing the number of minority class instances, giving the model more data to learn from the underrepresented class. Next comes ENN, a data-cleaning method that focuses on removing noisy or misclassified samples from both the majority and minority classes [23]. It works by examining the nearest neighbors of each instance (usually k=3). If the class label of the instance does not match the majority of its neighbors' class labels, the instance is removed. This technique helps to eliminate cases that are likely to be noisy or close to the decision boundary, improving the clarity of the dataset and the model's ability to learn. The advantages of this technique include Improved balance: SMOTE helps by oversampling the minority class, balancing the dataset, and making it easier for models to learn patterns for both classes. Noise reduction: ENN removes noisy or marginal samples that may confuse the model, resulting in clearer decision boundaries. Improved performance: The combination of oversampling and data cleaning helps to reduce overfitting (from oversampling) and underfitting (from noise),

leading to better generalization and improved model performance on imbalanced datasets [24]. This technique generated the categories shown in Fig. 13 below, which are notable for their differences from previous techniques. The distribution of data in this technique is uneven, though not very large.

The SMOTE-ENN technique gave the best results achieved for the proposed model as shown in Fig. 14 below. It gave a prediction accuracy of 99%. It also gave the best F1-score for all classes. For all the above, this study concludes that the best SMOTE technique that can be used for classification is SMOTE-ENN.

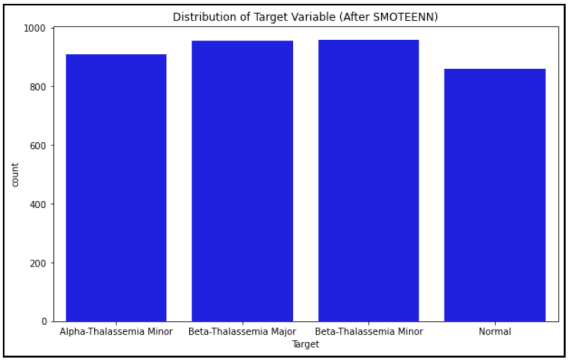

Fig. 13. Distribution of target variables in dataset (after SMOTE-ENN).

```
Classification Report :-
                        precision    recall  f1-score   support

Normal                       1.00      1.00      1.00       170
Beta-Thalassemia Minor       0.98      1.00      0.99       180
Beta-Thalassemia Major       0.97      0.98      0.98       180
Alpha-Thalassemia Minor      0.99      0.96      0.97       207

    accuracy                                     0.99       737
   macro avg                 0.99      0.99      0.99       737
weighted avg                 0.99      0.99      0.99       737


Confusion Matrix :-
[[170   0   0   0]
 [  0 180   0   0]
 [  0   0 177   3]
 [  0   3   5 199]]
```

Fig. 14. Results obtained from the random forest algorithm after applying the SMOTE-ENN technique.

3. Results

The proposed model in this study is mainly designed to make a scientific and practical comparison between the most important types of SMOTE on the one hand and on the other hand to study the prediction of thalassemia disease. The first stages in this model included reading the Iraqi local dataset and making the prediction of thalassemia disease using the random forest algorithm without

Table 1. Results obtained from the techniques used in the proposed model.

Technique	Thalassemia	Precision	Recall	F1-score	Accuracy
With Out SMOTE	Normal	1.00	0.99	1.00	0.99
	Beta-Thalassemia Minor	1.00	1.00	1.00	
	Beta-Thalassemia Major	0.96	1.00	0.98	
	Alpha-Thalassemia Minor	0.00	0.00	0.00	
SMOTE	Normal	1.00	1.00	1.00	0.98
	Beta-Thalassemia Minor	0.97	0.99	0.98	
	Beta-Thalassemia Major	0.95	0.99	0.97	
	Alpha-Thalassemia Minor	0.99	0.93	0.96	
SMOTENC	Normal	1.00	1.00	1.00	0.98
	Beta-Thalassemia Minor	1.00	0.97	0.99	
	Beta-Thalassemia Major	0.94	1.00	0.97	
	Alpha-Thalassemia Minor	0.98	0.95	0.96	
Borderline-SMOTE	Normal	1.00	1.00	1.00	0.98
	Beta-Thalassemia Minor	0.98	0.98	0.98	
	Beta-Thalassemia Major	0.98	0.98	0.98	
	Alpha-Thalassemia Minor	0.97	0.97	0.97	
SMOTE-Tomek	Normal	1.00	0.99	1.00	0.98
	Beta-Thalassemia Minor	0.99	0.99	0.99	
	Beta-Thalassemia Major	0.96	0.99	0.98	
	Alpha-Thalassemia Minor	0.97	0.95	0.96	
SMOTE-ENN	Normal	1.00	1.00	1.00	0.99
	Beta-Thalassemia Minor	0.98	1.00	0.99	
	Beta-Thalassemia Major	0.97	0.98	0.98	
	Alpha-Thalassemia Minor	0.99	0.96	0.97	

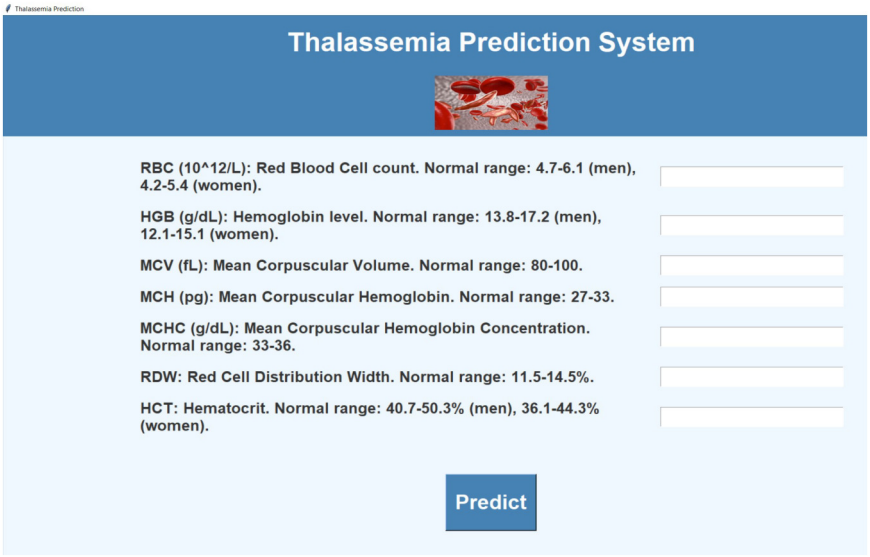

Fig. 15. The initial interface of the proposed model.

using SMOTE technology. Then came the other main steps represented by using multiple types of SMOTE to make a practical comparison of the results obtained from them. Below is Table 1, which shows the obtained results in detail.

From the table below, we can conclude the best technique for the best prediction result. The best accuracy is obtained in the first and last cases. However, the disadvantage of the first case - without using SMOTE - is the model's inability to predict any correct case of Alpha-Thalassemia Minor, which is considered the lowest category among the four categories. As for the last case - using SMOTE -ENN - in addition to the high prediction accuracy of 99%, the model was able to predict the case of Alpha-Thalassemia Minor with an accuracy of 97% according to the F1-score measure. Thus, the problem of bias in the model towards the higher categories at the expense of the lower categories was addressed, which created a balance in the categories of a single feature. SMOTE-ENN outperformed other techniques due to its combined approach of oversampling the minority class (via SMOTE) and removing noisy or borderline instances (via Edited Nearest Neighbors). This hybrid method enhances class balance while reducing overlapping between classes, leading to improved model generalization. Specifically, it proved effective in handling borderline cases, which were a major challenge in our dataset, as seen in its superior F1-score for the Alpha-Thalassemia Minor category. The interface required for the model's operation was designed as an integrated application, where the user enters the required test data and clicks the prediction button. Based on its training and the provided test results, the proposed model generates a prediction regarding the individual's likelihood of having thalassemia, as illustrated in Fig. 15 below.

4. Conclusion

Thalassemia is a disease that deserves to be highlighted. Predicting the occurrence of this disease for people at risk is one of the important things that helps to avoid and recover from it. On the other hand, researchers have always suffered from the problem of data imbalance and model bias towards one class at the expense of the other. SMOTE is one of the techniques that address this problem. As presented in this study, this technique has several types that have been discussed in detail. This study concluded that the best type of SMOTE that can be used and that gave the best prediction is SMOTE-ENN. This study is considered a basis and reference for researchers wishing to use one of the types of SMOTE. It is also a strong reference for software developers to use the model code to develop mobile and computer applications to help individuals prevent thalassemia.

Acknowledgment

Authors are thankful to the Al-Hussein Medical City in Karbala Governorate in Iraq, for supporting this work with the thalassemia dataset.

Author contributions

The first author programmed the model. Prepared and wrote the study. The second author obtained

the dataset and supported the first researcher with scientific advice.

Funding

The authors acknowledge that this study was not funded by any party.

Declarations

Conflict of Interest: The authors of this study declare that they are not aware of any financial interests that would conflict with this work or personal relationships that would interfere with the work performed.

References

1. A. Alsahli and M. Alsulmi, "Automatic detection of sand fouling levels in railway tracks using supervised machine learning: A case study from Saudi Arabian railway," *Arabian Journal for Science and Engineering*, vol. 48, no. 4, pp. 4925–4935, 2023.
2. J. David Weatheral and J. Clegg, *The Thalassaemia Syndromes*, 4th ed., Wiley-Blackwell, Oxford Book, 2001.
3. H. M. Alhuthali, *et al.*, "Molecular patterns of alpha-thalassemia in the kingdom of Saudi Arabia: Identification of prevalent genotypes and regions with high incidence," *Thrombosis Journal*, vol. 21, no. 1, p. 115, 2023.
4. W. Joint and W. H. Organization, "Management of birth defects and haemoglobin disorders: Report of a joint WHO-March of Dimes meeting, Geneva, Switzerland, 17–19 May 2006," In *Proc. Management of birth defects and haemoglobin disorders: report of a joint WHO-March of Dimes meeting*, Geneva, Switzerland, 17–19 May 2006, 2006.
5. A. Devanath, S. Akter, P. Karmaker, and A. Sattar, "Thalassemia prediction using machine learning approaches," In *Proc. 6th International Conference on Computing Methodologies and Communication (ICCMC)*, 2022: IEEE, pp. 1166–1174.
6. A. S. AlAgha, H. Faris, B. H. Hammo, and A.-Z. Ala'M, "Identifying β-thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine," *Artificial intelligence in medicine*, vol. 88, pp. 70–83, 2018.
7. M. Saleem, W. Aslam, M. I. U. Lali, H. T. Rauf, and E. A. Nasr, "Predicting Thalassemia using feature selection techniques: A comparative analysis," *Diagnostics*, vol. 13, no. 22, p. 3441, 2023.
8. J. V. Dacie, *Dacie and Lewis practical haematology*. Elsevier Health Sciences, 2006.
9. K. Doig and B. Zhang, "A methodical approach to interpreting the red blood cell parameters of the complete blood count," *American Society for Clinical Laboratory Science*, vol. 30, no. 3, pp. 173–185, 2017.
10. H. M. Merdas and A. H. Mousa, "Food sales prediction model using machine learning techniques," *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 13, no. 6, 2023.
11. A. J. Abdlmutalib and A. Abdelkarim, "Machine learning-based prediction of pore types in carbonate rocks using elastic properties," *Arabian Journal for Science and Engineering*, pp. 1–16, 2024.
12. R. Casarin, A. Facchinetti, D. Sorice, and S. Tonellato, "Decision trees and random forests," In *The Essentials of Machine Learning in Finance and Accounting*: Routledge, 2021, pp. 7–36.
13. R. Yacouby and D. Axman, "Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models," In *Proceedings of the first workshop on evaluation and comparison of NLP systems*, 2020, pp. 79–91.
14. K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged F1 and macro-averaged F1 scores," *Applied Intelligence*, vol. 52, no. 5, pp. 4961–4972, 2022.
15. D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Machine Learning*, vol. 113, no. 7, pp. 4903–4923, 2024.
16. S. Feng, J. Keung, P. Zhang, Y. Xiao, and M. Zhang, "The impact of the distance metric and measure on SMOTE-based techniques in software defect prediction," *Information and Software Technology*, vol. 142, p. 106742, 2022.
17. M. Mukherjee and M. Khushi, "SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features," *Applied system innovation*, vol. 4, no. 1, p. 18, 2021.
18. E. C. Gök and M. O. Olgun, "SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples," *Neural Computing and Applications*, vol. 33, no. 22, pp. 15693–15707, 2021.
19. N. A. Azhar, M. S. M. Pozi, A. M. Din, and A. Jatowt, "An investigation of smote based methods for imbalanced datasets with data complexity analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 6651–6672, 2022.
20. M. Revathi and D. Ramyachitra, "A modified borderline smote with noise reduction in imbalanced datasets," *Wireless Personal Communications*, vol. 121, no. 3, pp. 1659–1680, 2021.
21. K. Nugroho, A. R. Muslikh, S. W. Irianandа, and A. A. Ojugo, "Integrating SMOTE-Tomek and fusion learning with XGBoost meta-learner for robust diabetes recognition," *Journal of Future Artificial Intelligence and Technologies*, vol. 1, no. 1, 2024.
22. I. Papailia, "Predicting credit card fraud using machine learning techniques: Dealing with an imbalanced dataset," Vrije Universiteit Amsterdam, 2024.
23. N. Pramanick, S. Srivastava, J. Mathew, and M. Agarwal, "Enhanced IDS using BBA and SMOTE-ENN for imbalanced data for cybersecurity," *SN Computer Science*, vol. 5, no. 7, p. 875, 2024.
24. M. Kumari and N. Subbarao, "A hybrid resampling algorithms SMOTE and ENN based deep learning models for identification of Marburg virus inhibitors," *Future medicinal chemistry*, vol. 14, no. 10, pp. 701–715, 2022.