IMPROVING MACHINE LEARNING CLASSIFICATION MODELS FOR ANAEMIA TYPE PREDICTION BY OVERSAMPLING IMBALANCED COMPLETE BLOOD COUNT DATA WITH SMOTE-BASED ALGORITHMS

[a]LADISLAV VÉGH, [b]NORBERT ANNUŠ, [c]KRISZTINA CZAKÓOVÁ, [d]ONDREJ TAKÁČ

J. Selye University, Faculty of Economics and Informatics, Department of Informatics, Bratislavská cesta 3322, 945 01, Komárno, Slovakia
email: [a]veghl@ujs.sk, [b]annusn@ujs.sk, [c]czakoovak@ujs.sk, [d]takaco@ujs.sk

The paper was supported by the national project KEGA 014TTU-4/2024 "Intelligent Animation-Simulation Models, Tools, and Environments for Deep Learning."

Abstract: Computer-assisted disease diagnosis is cost-effective and time-saving, increasing accuracy and reducing the need for an additional workforce in medical decision-making. In our prior research, we trained, tested, and compared the accuracies of nine optimizable classification models to diagnose and predict eight anaemia types from Complete Blood Count (CBC) data. This study aimed to improve these classification models by oversampling the original imbalanced dataset with four algorithms related to the Synthetic Minority Over-sampling Technique (SMOTE). The results showed that the validation accuracy increased from 99.22% (Ensemble model) to 99.57% (Tree model), and most importantly, the False Discovery Rate (FDR) for the anaemia type with the highest FDR decreased from 23.1% to 1.5%.

Keywords: Anaemia Types, Classification, Complete Blood Count (CBC), Healthcare Analytics, Imbalanced Dataset, Machine Learning, Oversampling, SMOTE

1 Introduction

Artificial intelligence, deep learning, and machine learning are nowadays used in various fields [1–4] to recognise patterns in data, classify data and predict some outputs. Using modern information technologies in healthcare, driven by artificial intelligence, can improve medical diagnostics, reduce human errors and enhance timely detection [5]. Artificial intelligence, machine learning and deep learning techniques can be used in healthcare, for example, to predict cardiovascular disease [6,7] or diabetes [8,9] or to diagnose cancer [10]. Our research focuses on diagnosing and predicting eight anaemia types from Complete Blood Count (CBC) data. In prior research, we trained and tested nine optimisable classification models to predict anaemia types. Even though our best classification model reached 99.22% overall validation accuracy, the False Discovery Rate (FDR) was very high, 23.1%, for the Leukaemia with the thrombocytopenia category [11]. In the research presented in this paper, our primary goal was not only to improve the overall validation accuracy but, most importantly, to decrease the FDR for all the categories and, with that, to develop a more reliable machine learning classification model for anaemia diagnosis. To reach our goal, first of all, we oversampled the original imbalanced dataset using Synthetic Minority Over-sampling Techniques (SMOTE) [12–15]. Next, we trained and tested the machine learning classification models with the new datasets and compared the obtained results. Finally, we examined the classification models with the highest accuracies more deeply and developed a MATLAB app for predicting anaemia type from CBC data.

2 Related Research

Anaemia is a condition marked by inadequate red blood cells or haemoglobin, which decreases oxygen delivery to the body's tissues. It is estimated that around 40% of children aged 6 to 59 months, 37% of pregnant women, and 30% of women between the ages of 15 and 49 worldwide are impacted by anaemia [16]. Proper diagnosis and classification of anaemia are crucial for effective treatment. Historically, this process relies on CBC tests and manual evaluations by healthcare professionals, which can be time-consuming and vary due to subjective interpretations. However, recent advancements in data science and machine learning open up new avenues for enhancing the accuracy and efficiency of anaemia diagnosis [17].

Bahadure et al. [5] focused on identifying anaemia and its subtypes using deep learning and YOLO object detection algorithms to analyse microscopic images of blood samples. They highlighted the significance of quality metrics and feature extraction for ensuring precise diagnosis. Their research achieved an accuracy of 97.60% in identifying anaemia and its subtypes on 448×448 resolution images. Similarly, microscopic images of blood samples were used in research conducted by Dalvi and Vernekar [18]. First, they assessed 13 geometric features of each red blood cell; afterwards, they compared five ensemble methods to diagnose anaemia. The results showed that the best performance (overall accuracy: 92.122%, specificity: 95.082%, recall: 79.88%, precision: 80.134%) was reached with the Stacking ensemble method, which combined K-Nearest Neighbours (KNN) and Decision Tree as base learners and Naïve Bayes as stacking learner.

In a study by Airlangga [17] the researchers used CBC data to diagnose anaemia and compare the performance of various machine learning classification models. The results showed that the Decision Tree classifier achieved the highest balanced accuracy score of 94.17%, outperforming more complex ensemble methods. Aditya et al. [19] used CBC results as well to detect anaemia. They compared the accuracies of four machine learning models. The highest accuracy, 99.22%, was achieved with the Random Forest using the Staking CatBoost algorithm. Similarly, CBC results were used in a research by Faraj [20], where six classification algorithms were used on a dataset containing records of 180 women to diagnose anaemia. The highest accuracy, 97.78%, was reached with the Logistic Model Tree algorithm. Also, CBC data was used in a study by Pullakhandam and McRoy [21]. They reached 97% accuracy with six classification models: Logistic Regression, Random Forest, K-Nearest Neighbours (KNN), Gradient Boosting, and XGBoost. The analysis also showed that among all predictors, the low blood level of haemoglobin, the higher age, and the higher red blood cell distribution width were the most critical. Rahman et al. [22] also used the attributes of CBC data to diagnose anaemia. They used gender, age, haemoglobin, count of red blood cells (RBC), packed cell volume (PCV), mean corpuscular volume (MCV), mean corpuscular haemoglobin (MCH), and mean corpuscular haemoglobin concentration (MCHC) as input parameters. They compared the results of 11 machine learning algorithms; the highest accuracy, 95%, was reached with the Logistic Regression classification model. Vohra et al. [23] also used CBC data to diagnose anaemia and classify it into mild, moderate, and severe categories. Except for the original dataset, they used a feature-selected dataset, an oversampled dataset, and a feature-selected and oversampled dataset to train and test six machine learning classification models. For all models, both the hold-out experiment method and the 10-fold cross-validation method were used, and finally, the obtained results were compared. The highest accuracy, 99.35%, was reached with Multilayer Perceptron, using the hold-out experimental method and the dataset oversampled with SMOTE. Yıldız et al. [24] used hemogram data together with general information (such as age, sex, chronic diseases, and symptoms) to diagnose twelve different types of anaemia. They analysed the results of four machine learning algorithms trained and tested on the given dataset. The highest accuracy, 85.6%, was achieved using the Bagged Decision Trees classification model. Kovačević et al. [25] used the KNN machine learning algorithm to predict anaemia. They used gender, age, presence of chronic disease, iron, mean corpuscular volume (MCV), folic acid, haemoglobin, ferritin, erythrocyte count, and vitamin B12 as input parameters to diagnose and classify three types of anaemia with 98.4% accuracy. The study's results also showed that the most relevant parameter was the MCV.

3 Materials and Methods

In this research, we used a publicly available dataset containing CBC data downloaded from Kaggle [26]; and MATLAB R2024b software [27] for data exploration, oversampling, training and testing several machine learning classification models, and calculating the final models' feature importance.

3.1 Dataset

The original dataset [26] contained 1281 observations, 14 predictors from CBC data, and the diagnosis as a categorical target variable. The predictors were the following: the amount of haemoglobin (HGB), the number of platelets (PLT), the count of white blood cells (WBC), the count of red blood cells (RBC), the haematocrit test (HCT), the mean corpuscular volume (MCV), the mean corpuscular haemoglobin (MCH), the mean corpuscular haemoglobin concentration (MCHC), the variability in platelet size distribution in the blood (PDW), the procalcitonin test (PCT), the percent of lymphocytes (LYMp), the percent of neutrophils (NEUTp), the number of lymphocytes (LYMn), and the number of neutrophils (NEUTn). The categorical target variable of the dataset contained nine categories: one for healthy patients and eight for different types of anaemia. The anaemia types were the following: iron deficiency anaemia, leukaemia, leukaemia with thrombocytopenia, macrocytic anaemia, normocytic hypochromic anaemia, normocytic normochromic anaemia, other microcytic anaemia, and thrombocytopenia.

We divided the original dataset into a training set (90% of observations) and a test set (10% of observations). The training set was used to train classification models and to develop four new training sets using SMOTE-related algorithms (Fig. 1). The test set was used to test all our trained models on previously unseen data.

3.2 Data Exploration

Exploring datasets visually with many predictors can be challenging. Nevertheless, Principal Component Analysis (PCA) helps to decrease dimensionality and enhances interpretability while minimising the loss of information. This multivariate method is used to analyse datasets with inter-correlated quantitative dependent variables. The main goal of PCA is to extract important information from the dataset, which is then represented as a set of new uncorrelated variables that sequentially maximise variance. Furthermore, PCA uncovers patterns of similarity among the observations and variables, enabling these relationships to be illustrated as points on maps [28,29]. We used PCA to generate two principal components from 14 predictors and visualise these components in a 2D graph. This method allowed us to visually observe which target categories are similar or different from others.

3.3 Oversampling

An imbalanced dataset, where the distribution of classes is not uniform, can pose significant challenges in machine learning, particularly when certain categories have very few samples. As a result, classification algorithms struggle to predict these underrepresented categories accurately. A straightforward solution to this issue is to increase the number of records in the minority classes [30]. Traditionally, oversampling techniques involve duplicating samples from minority classes, such as Random Over-sampling (ROS) [31]. A widely used method is the Synthetic Minority Over-sampling Technique (SMOTE), which creates new synthetic samples by interpolating between existing minority-class samples [12]. This method has influenced various sampling techniques, including Borderline SMOTE [14], Safe-level SMOTE [15], and Adaptive Synthetic Sampling (ADASYN) [13]. In the research presented in this paper, we used four SMOTE-related algorithms implemented in MATLAB [32] to oversample the original training set, expanding every minority class in the original dataset to 130 samples (Fig. 1). We then used the original and the new training sets to train several classification models and analyse the results.

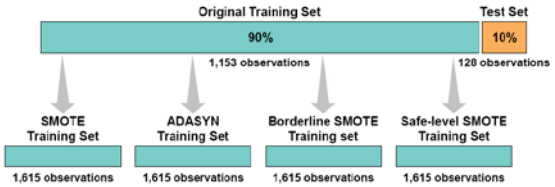

Figure 1: The training sets and the test set used in the research

3.4 Classification

Classification is a supervised machine learning technique that categorises data into predefined classes. In our research, we used several classification models to diagnose and predict eight types of anaemia. To accomplish this, we used the Classification Learner app [33] available in MATLAB R2004b software [27]. The models we employed include Ensemble, Tree, Support Vector Machine (SVM), Efficient Linear, Neural Network, Kernel, K-Nearest Neighbours (KNN), Naïve Bayes, and Discriminant analysis. We trained these models using the original and four oversampled training sets. To optimise the hyperparameters of each model, we applied Bayesian optimisation. Throughout the training process, we implemented 10-fold cross-validation to mitigate the risk of overfitting. Finally, we evaluated the performance of all models using an unseen test set.

3.5 Feature Importance

Although machine learning has significant potential to enhance products and services by quickly and accurately predicting outcomes from data, computers typically do not provide explanations for their predictions. To interpret these machine learning models, model-agnostic methods can be employed [34,35]. In this study, we calculated the permutation feature importance and Shapley importance for the two best classification models to understand how they operate.

4 Results

This section presents the study's findings, highlighting insights gained from data analysis and model evaluation. First, the data exploration results are summarised, including visual representations of the original and an oversampled training set projected onto two principal components. Next, the performance of various classification models is compared to evaluate their predictive accuracy. The analysis then identifies the most important features influencing predictions in the top two models. Finally, a MATLAB application is introduced as a practical tool for predicting different types of anaemia, demonstrating how the developed models can be applied in a user-friendly environment.

4.1 Data Exploration

First, we visualised the number of observations in the original training set for each target category. As we can see on the bar chart (Fig. 2), some target categories contain only a few observations. Using such an imbalanced dataset for training classification models might result in a high False Discovery Rate (FDR) for these categories, making the trained models unreliable for predicting them. To address this issue, we decided to generate oversampled training sets using SMOTE-based algorithms. We aimed to improve the classification models' accuracies and reduce the FDR for the most problematic categories.

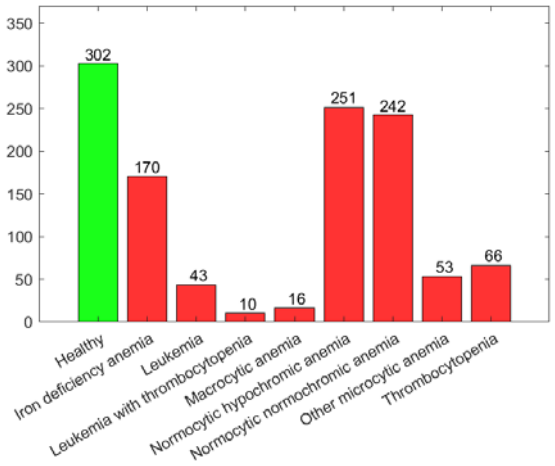

Figure 2: Number of observations by diagnostic categories in the original training set (N=1,153)

The dataset included 14 predictors, making visualisation only possible by using a technique to reduce its dimensions. Therefore, we opted for PCA to represent the target classes visually in a 2D graph. Fig. 3 displays the map of the two principal components for each category. Some classes are more distinguishable, while others, particularly the minority classes, are not visible in the graph.

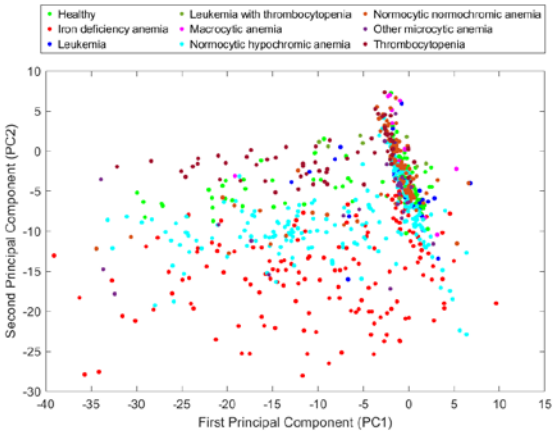

Figure 3: Original training set (N=1,153) projected onto two principal components

We oversampled the original training set by increasing the number of samples in the minority classes to 130 for five categories using four SMOTE-based algorithms [32]. The oversampled categories are Leukemia, Leukemia with thrombocytopenia, Macrocytic anaemia, Other microcytic anaemia, and Thrombocytopenia. While the original training set contained 1,153 observations, each of the oversampled training sets contained 1,615 samples. Fig. 4. shows the training set oversampled with SMOTE [12] projected onto two principal components. When comparing Fig. 3 to Fig. 4, we can observe the impact of SMOTE on the class distribution within the training set, emphasising its role in addressing class imbalance. As a result, more categories are distinguishable on the map; however, some observations from different categories remain close to one another.

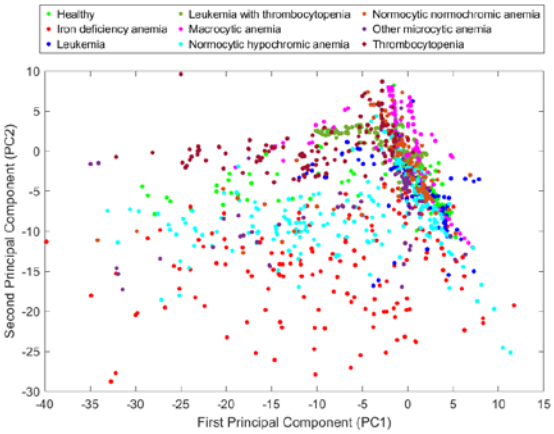

Figure 4: Oversampled (SMOTE) training set (N=1,615) projected onto two principal components

4.2 Comparison of Classification Models

After data exploration, we trained nine optimisable classification models first with the original training and then with the four oversampled training sets. During the training, 10-fold cross-validation was used. Tab. 1 summarises the validation accuracies of the trained models for each training set. As we can see in the table, the highest validation accuracies, 99.57% and 99.50%, were reached with Tree and Ensemble classification models that were trained on the training set oversampled with the standard SMOTE algorithm.

Tab. 1: Validation accuracies (%) of classification models on original (N=1,153) vs. four oversampled (N=1,615) training sets

Model	Training Set				
	Original	SMOTE	ADASYN	Borderline SMOTE	Safe-level SMOTE
Ensemble	99.22	99.50	99.44	99.44	99.38
Tree	99.05	99.57	99.20	99.26	99.32
SVM	91.76	94.61	94.67	94.06	94.67
Efficient Linear	89.51	92.38	92.69	92.57	92.63
Neural Network	88.03	94.55	93.37	94.67	82.54
Kernel	81.35	87.55	87.31	86.87	88.05
KNN	75.80	85.20	84.27	86.50	85.14
Naïve Bayes	67.30	71.02	72.76	73.13	73.19
Discriminant	54.38	59.94	60.50	60.06	58.02

After completing the training, we evaluated each classification model using an unseen test set. The test accuracies obtained are summarised in Tab. 2. For our best models, which achieved the highest validation accuracies, the test accuracy reached 100%, meaning they correctly classified every observation in the test set.

Tab. 2: Test accuracies (%) of trained classification models

Model	Training Set				
	Original	SMOTE	ADASYN	Borderline SMOTE	Safe-level SMOTE
Ensemble	100.00	100.00	100.00	100.00	98.44
Tree	100.00	100.00	100.00	100.00	100.00
SVM	91.41	92.97	92.19	95.31	94.53
Efficient Linear	94.53	93.75	92.97	93.75	92.97
Neural Network	89.06	93.75	91.41	96.09	84.38
Kernel	82.81	82.03	84.38	81.25	84.38
KNN	77.34	74.22	75.78	79.69	71.09
Naïve Bayes	67.19	69.53	67.19	58.59	70.31
Discriminant	52.34	53.12	54.69	53.91	53.91

In the following part of this paper, we will focus solely on our two best models: the Tree and Ensemble classification models. We will analyse their results in greater detail. Although these models achieved impressive overall validation and test accuracies, we still do not know how they classify observations within each category. Therefore, we should examine their confusion matrices. Fig. 5 displays the validation confusion matrices of the Ensemble and Tree models. The figure indicates that only one or two observations were misclassified in most categories.

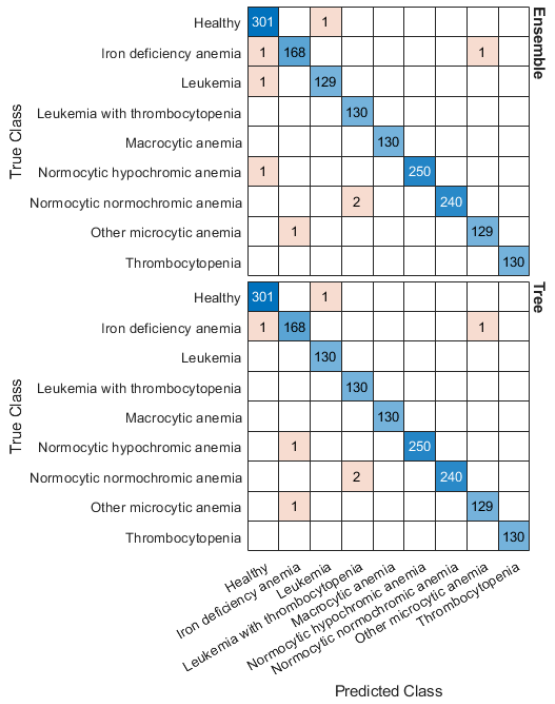

Figure 5: Validation confusion matrices of Ensemble and Tree classification models trained on the oversampled SMOTE training set

Fig. 6 shows the test confusion matrices for our best models. Both machine learning models correctly classified all observations in every category in the test set.

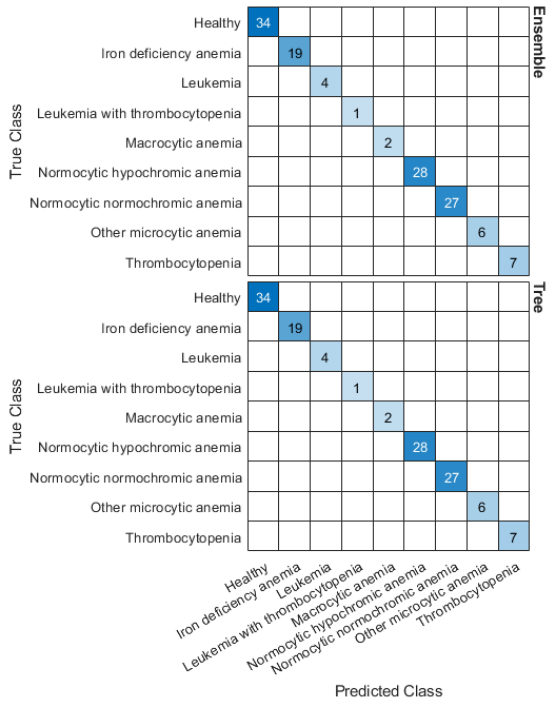

Figure 6: Test confusion matrices of Ensemble and Tree classification models trained on the oversampled SMOTE training set

The False Negative Rate (FNR) refers to the percentage of true positive cases a model mistakenly identifies as negative. It captures the rate of missed detections in the model's predictions. Tab. 3 summarises the FNR from the validation confusion matrices for our best models, which were trained on original and oversampled training sets. The data in the table shows that when we used the original training set with the Ensemble model, the categories with higher FNR were Macrocytic anaemia (FNR: 6.2%) and Other microcytic anaemia (FNR: 5.7%). In the case of the Tree model trained on the original training set, the categories that posed problems included Leukemia with thrombocytopenia (FNR: 10.0%), Macrocytic anaemia (FNR: 6.2%), and Other microcytic anaemia (FNR: 7.5%). However, when we applied the oversampled SMOTE training set, the FNR for all problematic classes dropped below 1.0% for both classification models.

Tab. 3: False negative rates (%) in validation confusion matrices of Ensemble and Tree models trained on original vs. oversampled training sets

Training Set, Model	**True Class**								
	Healthy	**Iron deficiency anaemia**	**Leukemia**	**Leukemia with thrombocytopenia**	**Macrocytic anaemia**	**Normocytic hypochromic anaemia**	**Normocytic normochromic anaemia**	**Other microcytic anaemia**	**Thrombocytopenia**
Original, Ensemble	0.3	0.6	0.0	0.0	6.2	0.4	0.8	5.7	0.0
Original, Tree	0.3	0.6	0.0	10.0	6.2	0.4	0.8	7.5	0.0
SMOTE, Ensemble	0.3	1.2	0.8	0.0	0.0	0.4	0.8	0.8	0.0
SMOTE, Tree	0.3	1.2	0.0	0.0	0.0	0.4	0.8	0.8	0.0
ADASYN, Ensemble	0.3	0.6	0.0	0.0	0.0	0.4	0.8	3.1	0.0
ADASYN, Tree	0.7	1.2	0.0	0.0	0.0	0.4	0.8	3.1	1.5
Borderline SMOTE, Ensemble	0.3	1.8	0.0	0.0	0.8	0.8	0.8	0.0	0.0
Borderline SMOTE, Tree	0.3	1.2	0.0	0.0	0.0	0.4	0.8	3.1	1.5
Safe-level SMOTE, Ensemble	0.0	0.6	0.8	0.0	0.8	1.2	0.8	0.8	0.8
Safe-level SMOTE. Tree	0.3	1.8	0.0	0.0	0.0	0.4	0.8	2.3	0.8

The False Discovery Rate (FDR) refers to the proportion of positive predictions made by a model that turn out to be false positives. It measures the likelihood of a positive prediction being incorrect. Tab. 4 presents the FDR from the validation confusion matrices of our best models, which were trained on original versus oversampled training sets. For both classification models trained on the original dataset, the most challenging category was Leukemia with thrombocytopenia, with an FDR of 23.1% in the Ensemble model and 18.2% in the Tree model. However, after applying the oversampled SMOTE training set, the FDR of the problematic class significantly decreased to 1.5% in both models.

Tab. 4: False discovery rates (%) in validation confusion matrices of Ensemble and Tree models trained on original vs. oversampled training sets

Training Set, Model	**Predicted Class**								
	Healthy	**Iron deficiency anaemia**	**Leukemia**	**Leukemia with thrombocytopenia**	**Macrocytic anaemia**	**Normocytic hypochromic anaemia**	**Normocytic normochromic anaemia**	**Other microcytic anaemia**	**Thrombocytopenia**
Original, Ensemble	0.7	0.6	2.3	23.1	0.0	0.4	0.0	0.0	1.5
Original, Tree	0.3	1.2	2.3	18.2	0.0	0.4	1.2	0.0	1.5
SMOTE, Ensemble	1.0	0.6	0.8	1.5	0.0	0.0	0.0	0.8	0.0
SMOTE, Tree	0.3	1.2	0.8	1.5	0.0	0.0	0.0	0.8	0.0
ADASYN, Ensemble	1.0	0.0	0.8	2.3	0.0	0.0	0.0	0.0	1.5
ADASYN, Tree	1.0	0.6	0.8	3.0	0.0	0.0	0.0	1.6	1.5

Borderline SMOTE, Ensemble	0.7	0.0	0.8	1.5	0.0	0.4	0.4	1.5	0.0
Borderline SMOTE, Tree	0.3	1.2	0.8	1.5	0.0	0.0	0.4	3.1	0.8
Safe-level SMOTE, Ensemble	1.0	0.6	0.0	1.5	0.0	0.4	0.8	0.0	0.8
Safe-level SMOTE. Tree	0.3	1.8	0.8	0.8	0.0	0.0	0.8	1.6	0.8

Since all the machine learning models we used are optimisable, we aimed to determine the optimal hyperparameters for our best classification models, specifically the Ensemble and Tree models. Tab. 5 displays the bestpoint hyperparameters for the Ensemble model. With these parameters, we observed a minimum classification error of 0.0037086.

Tab. 5: Bestpoint hyperparameters of the Ensemble model trained on the oversampled SMOTE training set

Hyperparameter	Value
Ensemble method:	Bag
Number of learners:	27
Maximum number of splits:	350
Number of predictors to sample:	7

Tab. 6 presents the optimal hyperparameters for the Tree classification model. The minimum classification error observed with these hyperparameters was 0.0043136.

Tab. 6: Bestpoint hyperparameters of the Tree model trained on the oversampled SMOTE training set

Hyperparameter	Value
Maximum number of splits:	96
Split criterion:	Maximum deviance reduction

4.3 Identifying Key Features in Tree and Ensemble Models' Predictions

Although we achieved high accuracy, low False Negative Rates (FNR), and low False Discovery Rates (FDR) with the Ensemble and Tree classification models when trained on the oversampled SMOTE training set, we still do not fully understand how these models operate. To gain deeper insights into the selected machine learning models and identify which of the 14 predictors are most significant in their predictions, we calculated both permutation feature importance and Shapley importance for each feature.

As a result of calculating the permutation feature importance, Fig. 7 illustrates the mean importance of each predictor for the Ensemble classification model, while Fig. 8 illustrates the mean importance of each predictor for the Tree classification model. The charts indicate that the most significant predictors in both machine learning models were the following: the amount of haemoglobin (HGB), the mean corpuscular volume (MCV), the platelet count (PLT), the white blood cell count (WBC), the mean corpuscular haemoglobin concentration (MCHC), the mean corpuscular haemoglobin (MCH), and the haematocrit test (HCT).

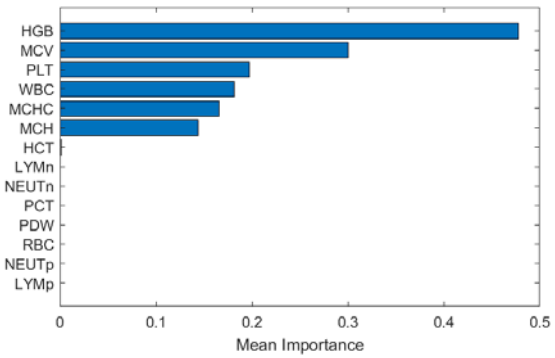

Figure 7: Mean importance per predictor of the Ensemble model trained on the oversampled SMOTE training set

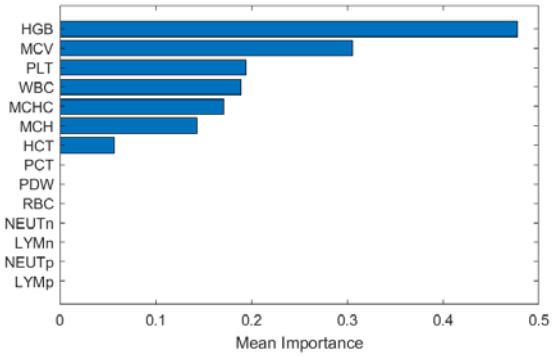

Figure 8: Mean importance per predictor of the Tree model trained on the oversampled SMOTE training set

Fig. 9 and Fig. 10 illustrate the Shapley importance values for each predictor variable in the chosen Ensemble and Tree classification models. While the significance of these predictors closely aligns with the results obtained through permutation feature importance, the Shapley graphs offer a more nuanced understanding of each predictor's contribution to the model's outputs.

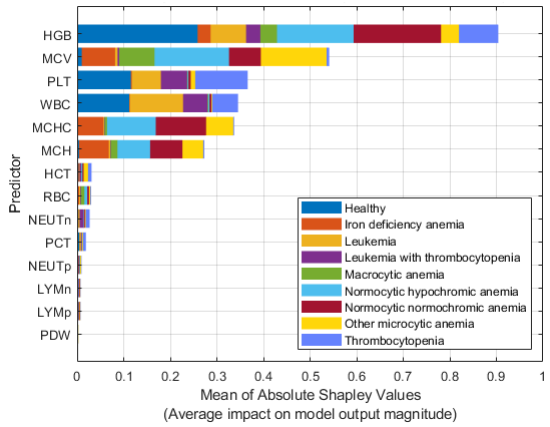

Figure 9: Shapley importance per predictor of the Ensemble model trained on the oversampled SMOTE training set

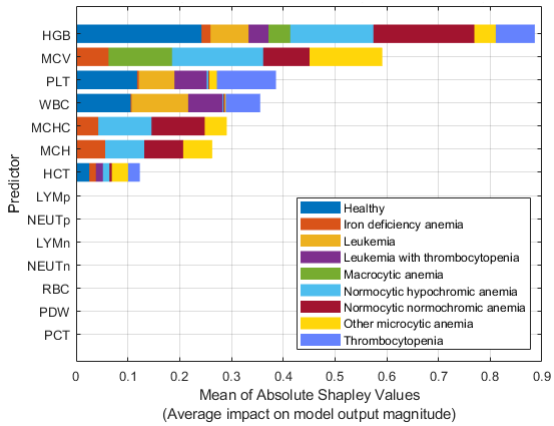

Figure 10: Shapley importance per predictor of the Tree model trained on the oversampled SMOTE training set

4.4 MATLAB App for Anaemia Types Prediction

As a practical outcome of our research, we developed a MATLAB app (Fig. 11) for diagnosing anaemia and predicting its subtype based on Complete Blood Count (CBC) data. Users can enter values for 14 predictors derived from the CBC data, and the app will calculate and visualise the likelihood of each type of anaemia using our Ensemble and Tree classification

models. The MATLAB app and the exported Ensemble and Tree classification models are available on GitHub [36].

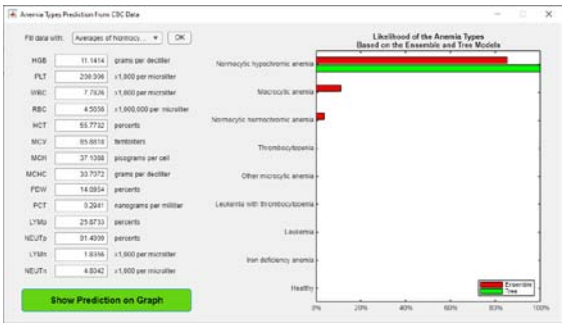

Figure 11: MATLAB App for Anaemia Types Prediction

5 Discussion

The results of this study highlight the potential of SMOTE-based oversampling techniques in addressing the challenges posed by imbalanced datasets in machine learning classification models. By applying these methods, we observed significant improvements in the performance metrics of the models used for anaemia type prediction. The validation and test accuracies of the models trained on the oversampled datasets outperformed those trained on the original dataset, with the Tree and Ensemble models achieving remarkable validation accuracies of 99.57% and 99.50% and test accuracies of 100%. Compared to previous studies [17,19–25], which achieved accuracies ranging from 85.6% to 99.35% using various machine learning algorithms; our approach consistently delivered higher accuracy. This is primarily attributed to the enhanced representation of minority classes through oversampling, reducing False Negative Rates (FNR) and False Discovery Rates (FDR) across all categories. For instance, the FDR for the previously problematic Leukemia with thrombocytopenia category decreased from 23.1% to 1.5% when using the oversampled SMOTE dataset. The findings also align with other research efforts, emphasising the importance of balancing datasets for improved classification outcomes [17,21,23,25]. However, our study extends this knowledge by evaluating the impact of multiple SMOTE variations, including ADASYN, Borderline SMOTE, and Safe-level SMOTE, on model performance. Among these, the standard SMOTE algorithm consistently yielded the best results regarding accuracy and reliability.

The feature importance analysis provided valuable insights into the predictors driving model decisions. Features such as haemoglobin levels (HGB), mean corpuscular volume (MCV), and platelet count (PLT) emerged as critical indicators for differentiating anaemia types. This aligns with clinical understandings, reinforcing the relevance of the developed models [37,38].

While the results are promising, this study has limitations. The dataset's reliance on CBC data alone may not capture the full clinical complexity of anaemia. Further validation on larger, more diverse datasets is necessary to generalise these findings to broader populations.

6 Conclusion

This research demonstrates the efficacy of applying SMOTE-based oversampling techniques to improve the performance of machine learning models for anaemia type prediction. By addressing class imbalances, the developed Tree and Ensemble machine learning models achieved high accuracy (99.57% and 99.50%), reduced error rates, and provided reliable predictions, even for minority classes. The findings underscore the importance of data preprocessing in healthcare analytics, particularly for imbalanced datasets. A MATLAB application, offering a user-friendly tool for real-world implementation, further showcased the practical utility of the developed models.

Future research should explore integrating additional clinical data and testing these methods on larger datasets to enhance their applicability and robustness. Nevertheless, the approach presented here sets a solid foundation for leveraging machine learning to advance medical diagnostics.

Literature:

1. Udvaros, J., Forman, N.: *Artificial Intelligence and Education 4.0*. Valencia, Spain; 2023. pp. 6309–6317. https://doi.org/10.21125/inted.2023.1670
2. Szénási, S., Légrádi, G., Vígh, B.: *Machine Learning-Assisted Approach for Optimizing Step Size of Hill Climbing Algorithm*. 2024 IEEE 18th International Symposium on Applied Computational Intelligence and Informatics (SACI). Timisoara, Romania: IEEE; 2024. pp. 000425–000430. https://doi.org/10.1109/SACI60582.2024.10619891
3. Annuš, N.: *Usability of Artificial Intelligence to Create Predictive Models in Education*. Palma, Spain; 2023. pp. 5061–5065. https://doi.org/10.21125/edulearn.2023.1328
4. Végh, L., Czakóová, K., Takáč, O.: *Comparing Machine Learning Classification Models on a Loan Approval Prediction Dataset*. International Journal of Advanced Natural Sciences and Engineering Researches. 2023, 7(9), pp. 98–103. https://doi.org/10.59287/ijanser.1516
5. Bahadure, N. B., Khomane, R., Nittala, A.: *Anemia Detection and Classification from Blood Samples Using Data Analysis and Deep Learning*. Automatika. 2024, 65(3), pp. 1163–1176. https://doi.org/10.1080/00051144.2024.2352317
6. Subramani, S., Varshney, N., Anand, M. V., Soudagar, M. E. M., Al-keridis, L. A., Upadhyay, T. K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K., Rohini, K.: *Cardiovascular diseases prediction by machine learning incorporation with deep learning*. https://doi.org/10.3389/fmed.2023.1150933
7. Végh, L., Takáč, O., Czakóová, K., Dancsa, D., Nagy, M.: *Comparative Analysis of Machine Learning Classification Models in Predicting Cardiovascular Disease*. International Journal of Advanced Natural Sciences and Engineering Researches. 2024, 8(6), pp. 23–31.
8. Mujumdar, A., Vaidehi, V.: *Diabetes Prediction using Machine Learning Algorithms*. Procedia Computer Science. 2019, 165, pp. 292–299. https://doi.org/10.1016/j.procs.2020.01.047
9. Tasin, I., Nabil, T. U., Islam, S., Khan, R.: *Diabetes prediction using machine learning and explainable AI techniques*. Healthc Technol Lett. 2022, 10(1–2), pp. 1–10. https://doi.org/10.1049/htl2.12039
10. Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., Waddell, N.: *Deep learning in cancer diagnosis, prognosis and treatment selection*. Genome Medicine. 2021, 13(1), pp. 152. https://doi.org/10.1186/s13073-021-00968-x
11. Végh, L., Takáč, O., Czakóová, K., Dancsa, D., Nagy, M.: *Evaluating Optimizable Machine Learning Models for Anemia Type Prediction from Complete Blood Count Data*. International Journal of Advanced Natural Sciences and Engineering Researches. 2024, 8(7), pp. 108–119.
12. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.: *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research. 2002, 16, pp. 321–357. https://doi.org/10.1613/jair.953
13. He, H., Bai, Y., Garcia, E. A., Li, S.: *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). 2008. pp. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969
14. Han, H., Wang, W.-Y., Mao, B.-H.: *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning*. https://doi.org/10.1007/11538059_91
15. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: *Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem*. https://doi.org/10.1007/978-3-642-01307-2_43
16. *Anaemia*. https://www.who.int/news-room/fact-sheets/detail/ANAEMIA
17. Airlangga, G.: *Leveraging Machine Learning for Accurate Anemia Diagnosis Using Complete Blood Count Data*.

Indonesian Journal of Artificial Intelligence and Data Mining. 2024, 7(2), pp. 318–326. https://doi.org/10.24014/ijaidm.v7i2.29869
18. Dalvi, P. T., Vernekar, N.: *Anemia Detection Using Ensemble Learning Techniques and Statistical Models*. 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). Bangalore, India: IEEE; 2016. pp. 1747–1751. https://doi.org/10.1109/RTEICT.2016.7808133
19. Aditya, M. R., Sutanto, T., Budiman, H., Ridha, M. R. N., Syapotro, U., Azijah, N.: *Machine Learning Models for Classification of Anemia from CBC Results: Random Forest, SVM, and Logistic Regression*. Journal of Data Science. 2024. https://iuojs.intimal.edu.my/index.php/jods/article/view/589
20. Faraj, S. M.: *Performance Evaluation of Machine Learning Algorithms for Predictive Classification of Anemia Data*. 2024.
21. Pullakhandam, S., McRoy, S.: *Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning*. BioMedInformatics. 2024, 4(1), pp. 661–672. https://doi.org/10.3390/biomedinformatics4010036
22. Rahman, Md. M., Mojumdar, M. U., Shifa, H. A., Chakraborty, N. R., Stenin, N. P., Hasan, Md. A.: *Anemia Disease Prediction using Machine Learning Techniques and Performance Analysis*. 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom). 2024. pp. 1276–1282. https://doi.org/10.23919/INDIACom61295.2024.10498962
23. Vohra, R., Hussain, A., Dudyala, A. K., Pahareeya, J., Khan, W.: *Multi-Class Classification Algorithms for the Diagnosis of Anemia in an Outpatient Clinical Setting*. PLoS One. 2022, 17(7), pp. e0269685. https://doi.org/10.1371/journal.pone.0269685
24. Karagül Yıldız, T., Yurtay, N., Öneç, B.: *Classifying Anemia Types Using Artificial Learning Methods*. Engineering Science and Technology, an International Journal. 2021, 24(1), pp. 50–70. https://doi.org/10.1016/j.jestch.2020.12.003
25. Kovacevic, A., Lakota, A., Kuka, L., Becic, E., Smajovic, A., Pokvic, L. G.: *Application of Artificial Intelligence in Diagnosis and Classification of Anemia*. 2022 11th Mediterranean Conference on Embedded Computing (MECO). Budva, Montenegro: IEEE; 2022. pp. 1–4. https://doi.org/10.1109/MECO55406.2022.9797180
26. *Anemia Types Classification*. https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification
27. *MATLAB*. https://www.mathworks.com/products/matlab.html
28. Abdi, H., Williams, L. J.: *Principal component analysis*. WIREs Computational Statistics. 2010, 2(4), pp. 433–459. https://doi.org/10.1002/wics.101
29. Jolliffe, I. T., Cadima, J.: *Principal component analysis: a review and recent developments.* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2016, 374(2065), pp. 20150202. https://doi.org/10.1098/rsta.2015.0202
30. Mohammed, R., Rawashdeh, J., Abdullah, M.: *Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results*. 2020 11th International Conference on Information and Communication Systems (ICICS). 2020. pp. 243–248. https://doi.org/10.1109/ICICS49469.2020.239556
31. Viloria, A., Pineda Lezama, O. B., Mercado-Caruzo, N.: *Unbalanced data processing using oversampling: Machine Learning*. Procedia Computer Science. 2020, 175, pp. 108–113. https://doi.org/10.1016/j.procs.2020.07.018
32. Michio, I.: *Oversampling Imbalanced Data: SMOTE related algorithms*. GitHub; 2024. https://github.com/minoue-xx/Oversampling-Imbalanced-Data/releases/tag/1.0.2
33. *Train models to classify data using supervised machine learning - MATLAB*. https://www.mathworks.com/help/stats/classificationlearner-app.html
34. Molnar, C.: *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/
35. Lundberg, S. M., Lee, S.-I.: *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
36. Végh, L.: *MATLAB App for Anemia Types Prediction from CBC Data*. GitHub; 2024. https://github.com/veghl/anemia/
37. Beyan, C., Kaptan, K., Beyan, E., Turan, M.: *The Platelet Count/Mean Corpuscular Hemoglobin Ratio Distinguishes Combined Iron and Vitamin B12 Deficiency from Uncomplicated Iron Deficiency*. International Journal of Hematology. 2005, 81(4), pp. 301–303. https://doi.org/10.1532/IJH97.E0311
38. Lin, H., Zhan, B., Shi, X., Feng, D., Tao, S., Wo, M., Fei, X., Wang, W., Yu, Y.: *The mean reticulocyte volume is a valuable index in early diagnosis of cancer-related anemia*. https://peerj.com/articles/17063

Primary Paper Section: I

Secondary Paper Section: IN, FD