An Efficient Algorithm for Diseases Classification based on Hemogram Blood Test Samples

Phuoc-Hai Huynh[1,2][0000-0001-8348-9267], Ngoc-Minh Nguyen[1,2][XXXX-XXXX-XXXX-XXXX], Thanh-Nghi Doan[1,2][0000-0001-8348-9267]

[1] Faculty of Information Technology, An Giang University, An Giang, Vietnam
[2] 2Vietnam National University Ho Chi Minh City, Vietnam

Abstract.

The growth of hospital information systems in recent years has resulted in a vast amount of medical data that must be studied to improve the quality and efficiency of medical services. Machine learning is a technology that can automatically and accurately analyze medical data, making it useful for disease diagnosis and treatment. In this paper, our goal is to compare different classification algorithms and determine the most effective one for hematological data. We propose an efficient algorithm for disease classification based on hemogram blood test samples using the random forest algorithm and the synthetic minority oversampling technique (SMOTE). Numerical test results on real hematological data from a local hospital show that the proposed method got an accuracy of up to 97.75% and an AUC of up to 98.65%, which is better than the best classification models currently available.

Keywords: machine learning, classification, random forest, SMOTE, hemogram blood test, medical informatics

1 Introduction

In recent years, the robust development of hospital information systems has led to a large volume of medical data that needs to be mined in order to increase the quality and efficiency of medical services. Medical data types are categories of information that are collected, stored, and analyzed in the context of healthcare. Examples of medical data types include demographic data, clinical data, laboratory data, imaging data, and genomic data. Hematological data comprises numerical or categorical values that describe the characteristics and functions of blood cells and their components. These values are obtained from hemogram blood test samples and are crucial clinical information used to diagnose, classify, monitor, and treat various diseases. Hemogram blood test samples provide valuable data for diagnosing and treating diseases, especially when many diseases have similar symptoms. Machine learning is a technology that can automatically and accurately analyze hemogram data [1]. These models are trained on a large dataset of patient blood test results and diagnoses, learning to distinguish between diseases based on the features of the blood test results. Additionally, machine learning models can detect diseases early, even when they are asymptomatic. This can improve treatment outcomes and reduce mortality rates [2].

Nowadays, more studies are being conducted to personalize treatment plans for individual patients . Machine learning models can predict how patients will respond to different treatments, helping doctors choose the best treatment for each patient. By using machine learning, doctors can make faster, more accurate, and more efficient diagnoses, improving treatment outcomes for patients and reducing healthcare costs [2]. There have been numerous studies on applying machine learning models for classifying diseases in biomedical data [3], [4]. However, there has been a lack of research on hematological data. In literature, many research papers are available which mainly focused on predicting diseases from hemogram data sets using data mining techniques. In [5], the authors presented a novel benchmark data set for blood diseases detection and applies several classical machine learning algorithms to classify diseases based on blood test parameters. Random Forest is proposed in [6]. This paper compares different classification algorithms using hematological data and finds that is the most effective algorithm with an accuracy of 96.47%. In addition, some studies ([7] and [8]) show that hemogram blood test samples could perform the analysis of COVID-19. S. Vijayarani and S. Sudha [9] proposed a new clustering algorithm which is named as weight based k-means algorithm is developed for identifying the leukemia, inflammatory, bacterial or viral infection, HIV infection and pernicious anaemia diseases from the hemogram blood test samples data set.

In this manuscript, we present an approach for classifying diseases using hemogram blood test samples. Specifically, we have proposed random forest [10] based on the oversampling algorithm by SMOTE [11] to diagnose six diseases. This approach, which uses hemogram blood test samples as data input and combines specific features throughout the forest, is able to classify diseases with an accuracy up to 97.75%. We have collected, cleaned, preprocessed, and labelled an amount of 1766 samples from the An Giang province regional general hospital (An Giang, Vietnam) to be fed to our algorithm. However, the data collected is unbalanced that led to the classify algorithms are overfiting. So the SMOTE is suggested to address this issue. This combine approach improves the classification performance of models. Furthermore, we have used comparison analysis techniques and visualize to interpret our results and compare them with stage-of-the-art classify methods.

The remainder of this paper is organized as follows. Section 2 describes material and methods. Results and discussion are explained in Section 3. At Section 4, the conclusion is given.

2 Materials and Methods

The primary goal of this research is to analyze hemogram blood test samples in order to predict diseases. The experiment was carried out in four main steps: data collection, data preprocessing, data augmentation, and classification.

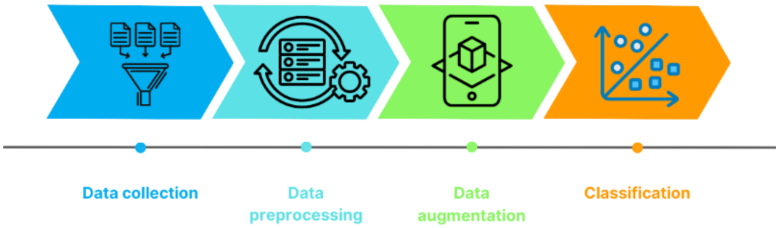

Figure 1: The overview experiment design

In the data collection step, the researchers obtained the necessary data from the laboratory information system (LIS) database of the An Giang Province Regional General Hospital. The database contained a total of 1,766 samples from patients.

After collecting the data, the next step involved data preprocessing. This step aimed to clean and prepare the data for further analysis. Outliers and missing values were removed from the dataset to ensure the quality of the data [12]. Additionally, the data was normalized to a common scale, which is often done to bring different features or variables into a similar range.

The third step involved data augmentation using the SMOTE algorithm. This technique is commonly used to address class imbalance issues in machine learning datasets. By using SMOTE, it was possible to generate new samples by interpolating between existing ones, resulting in a more accurate representation of the dataset's various classes.

Finally, the preprocessed and augmented data was used for classification. Multiple machine learning algorithms were employed for this task, including support vector machines (SVMs) [13], random forests (RF) [10], neural networks (ANNs), decision trees (DT) [14], bagging of decision trees (BA) [15], gradient boosting of decision trees (GBDTs) [16], and k-nearest neighbors (kNN) [17]. Each algorithm likely has its own strengths and weaknesses, and by comparing their performance, we can evaluate which models perform best for the given task based on the experimental results. To assess the models' performance, we used various evaluation metrics. These included classification accuracy (CA), precision (Prec), recall (Recall), F1 score, and area under the curve (AUC) [18].

2.1 Dataset and Preprocessing

In the preprocessing step of the dataset, we use many preprocessing data methods to ensure the quality and reliability of the data for further analysis. These technologies include eliminating irrelevant attributes and refilling missing values. Firstly, irrelevant attributes were eliminated from the dataset. This step involved identifying and removing any attributes that were deemed unnecessary or not relevant to the analysis. All private patient information is removed to ensure information security. The dataset was streamlined to focus on the essential features related to the complete blood count (CBC) measurements. The attributes characterize the Complete Blood Count (CBC) features as in Table 1.

Table 1. Characteristics of Complete Blood Count (CBC) feautures

ID	Name Feature	Decription
1	Age	
2	Sex	
3	WBC (cmm)	White Blood Cell
4	RBC (million/cmm)	Red Blood Cell
5	HGB (g/dl)	Hemoglobin
6	HCT (I/I)	Hemoglobin
7	MCV (ft)	Mean Cellular Volume
8	MCH (pg)	Mean Cellular Hemoglobin
9	MCHC (g/dl)	Hemoglobin Concentration
10	PLT (/Cmm)	Platelet Count
11	LYM 1	Lymphocyte Percentage
12	LYM 2	Lymphocyte Percentage
13	MPV	Mean Platelet Volume
14	RDW	Red Cell Distribution Width

We chose six common illness types for categorization in specific disease groups based on the ICD-10 codes. They include L02, D69.3, D17, J02, D50 and A94. The information of these disease code could search in website (https://icd.who.int/browse10/2010/en).

In order to handle missing values, we refill missing values in the dataset using the SMOTE algorithm. Missing values can occur due to various reasons, such as data entry errors or equipment malfunctions. It is important to handle missing values appropriately to avoid biased or inaccurate results. The SMOTE technique is used to fill in the missing values based on the available data. This step ensures that all samples have complete information for the selected attributes.

2.2 Methods

In this session, we describe the primary methods that we used to develop our efficient algorithm for disease classification based on hemogram blood test samples, including the SMOTE algorithm and Random Forest. In other methods, we use them to compare our approach.

SMOTE Algorithm

For the collected datasets, it is clear that our study faces the challenge of working with imbalanced datasets (Figure 2). This issue is popular for medical data. Most machine learning techniques have poor performance on the minority class. One way to solve this problem is to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information for the model.

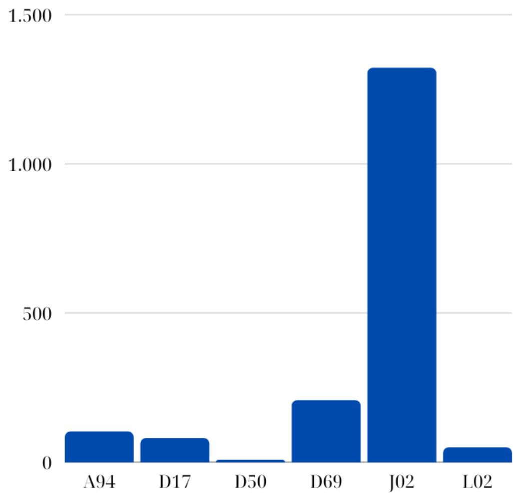

Figure 2: Data Distribution by Disease Label: Class Imbalance in the Dataset

SMOTE algorithms are a type of oversampling technique that are used to balance the classes in a dataset. We propose SMOTE to conjunct with classification methods to improve the accuracy of models as well as fill miss values. There are four steps in the SMOTE algorithm [11]:

Step 1: Determine the class in the dataset that has fewer instances compared to the majority class. This class is considered the minority class.

Step 2: Select a minority class instance at random to use as the foundation for creating fake samples.

Step 3: Compute the k-nearest neighbors: Calculate the k-nearest neighbors of the selected minority class instance. The value of k is a user-defined parameter that determines the number of neighbors to consider.

Step 4: For each case of a minority class, the algorithm makes a synthetic sample by picking at random one or more of its k closest neighbors. These samples are made by interpolating the feature values between the chosen instance and its neighbors.

Random Forest Classification

Random forest is an ensemble method that combines multiple decision trees to create a more accurate and robust classifier [10]. Each decision tree is trained on a random subset of the data and a random subset of the features. The final prediction is obtained by averaging the predictions of all the trees. Random forest can handle high-dimensional and noisy data and can also provide estimates of feature importance and error rates [19]. In addition, this algorithm is effective for handling imbalanced data because the random feature selection during tree construction helps in reducing bias towards the majority class [20]. In this study, we propose a random forest to classify diseases based on hemogram blood test samples. One of the advantages of a random forest is that it can capture the nonlinear and complex relationships between the features and the target variable. Moreover, random forest is easy to implement and interpret.

3 Evaluation and Discussion

We are interested in the classification performance of our proposal (using SMOTE and Random Forest) to classify diseases based on hemogram blood test samples. As a result, we present a comparison of the classification performance between our model and the top modern algorithms. In order to evaluate the effectiveness of classification tasks, we have implemented all algorithms in Python using the Scikit library [21].

3.1 Experiments setup

We use a 10-fold cross-validation protocol that remains the most widely used to evaluate its performance. The total classification metrics measure is used to evaluate the classification models. We train and evaluate the random forest classifier using cross-validation and various performance metrics, such as accuracy, precision, recall, F1-score, and ROC curve. In addition, we also compare our results with those of other machine learning methods, such as support vector machines (SVMs), random forests (RFs), neural networks (ANNs), decision trees (DTs), bagging of decision trees (BA-DTs), gradient boosting of decision trees (GB-DTs), and k-nearest neighbors (kNN). We show that our algorithm achieves high accuracy and low error rates in classifying different diseases based on hemogram blood test samples. All experiments are run on a machine running Linux Mint, with an Intel (R) Xeon (R) CPU at 3.07 GHz, 8 cores, and 8 GB of RAM.

3.2 Classification results

In order to demonstrate the effectiveness of the SMOTE and Random Forest, we conducted two experiments in a concise way, as follows:

In the first experiment, sevven classification approaches were used for direct classification without incorporating the SMOTE algorithm. The original hemogram dataset is imbalanced and is used for training and evaluating the classification models. The purpose of this experiment was to establish a baseline performance for comparison. Table 3 and Figure 3 show the classification results of this experiment.

Table 3 shows the performance of seven classifying models on a dataset of hemogram blood test samples. The best performing model was the RF model, which achieved a CA of 85.73%, a precision of 93.1%, a recall of 85.73%, an F1 score of 89.5%, and an AUC of 65.66%. The RF model performed better than the other models because it is able to learn complex relationships between the features and the diasease classes [22]. The other models performed well, but they did not perform as well as the RF model.

In addition, we also use AUC metric to evaluate classifying methods. The AUC metric is preferred for evaluating imbalanced data because it addresses the challenges posed by class imbalance and provides a comprehensive assessment of the model's discriminative ability, regardless of the class distribution [18]. However, the AUC metric results in Table 2 and Figure 3 show that these classifying models face imbance issue. While the accuracy of the models is very high, the AUC is not good.

Table 2. Comparison of various classifiers without using SMOTE.

Model	CA	Prec	Recall	F1	AUC
ANN	84.54	92.95	84.54	88.45	64.26
BAG	85.45	92.77	85.45	89.2	66.13
DT	76.67	75.32	76.67	75.98	65.99
GB	84.88	91.63	84.88	88.71	65.85
kNN	83.3	91.88	83.3	87.2	63.19
RF	**85.73**	93.71	**85.73**	**89.5**	65.66
SVM	84.54	93.94	84.54	88.85	63.63

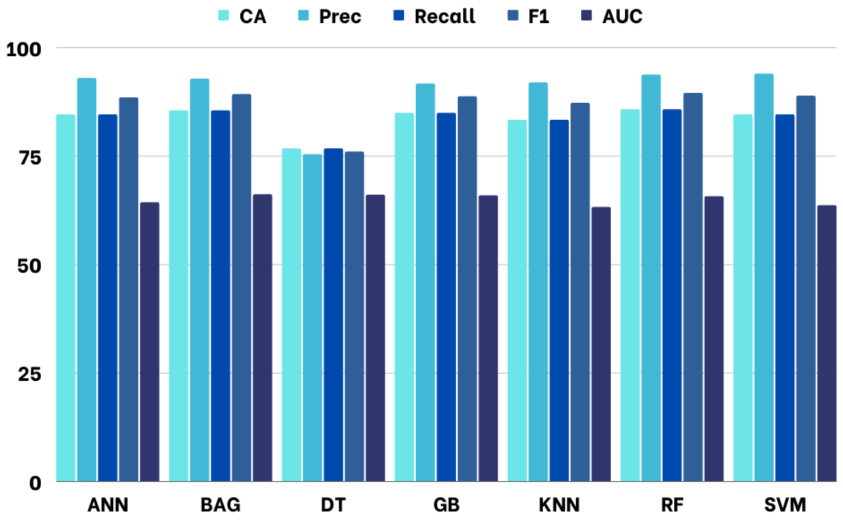

Figure 3. Classification results comparison without using SMOTE

In the second experiment, we employed the SMOTE algorithm to address the issue of class imbalance in the dataset. By oversampling the minority class using SMOTE, we aimed to rebalance the class distribution and create a more representative training set. The augmented dataset, consisting of both the original samples and the SMOTE-generated samples, was then used for training and evaluating the classification models in this experiment. Classification results are shown in Table 3 and Figure 4.

In Table 3, we present the comparison of various classifiers using the SMOTE algorithm for addressing class imbalances in the dataset. The classifiers evaluated in this study include SMOTE-ANN, SMOTE-BAG, SMOTE-DT, SMOTE-GB, SMOTE-KNN, SMOTE-RF, and SMOTE-SVM. Each model was trained and evaluated on the balanced dataset created using the SMOTE algorithm. The results demonstrate the performance of each classifier in terms of accuracy, precision, recall, F1-score, and AUC. Higher values indicate better performance across these metrics. The number results show that SMOTE-RF achieved a high classification accuracy of 97.75%, precision of 97.78%, recall of 97.75%, F1-score of 97.75%, and AUC of 98.65%.

Table 3. Comparison of various classifiers using SMOTE

Model	CA	Prec	Recall	F1	AUC
SMOTE-ANN	94.59	94.63	94.59	94.6	96.75
SMOTE-BAG	96.24	96.28	96.24	96.24	97.74
SMOTE-DT	91.22	91.3	91.22	91.24	94.73
SMOTE-GB	95.85	95.95	95.85	95.85	97.51
SMOTE-KNN	91.76	93.06	91.76	92	95.06
SMOTE-RF	**97.75**	**97.78**	**97.75**	**97.75**	**98.65**

SMOTE-SVM	94.55	94.75	94.55	94.58	96.73

In addition, by comparing the performance metrics of different classifiers, this table provides insights into their effectiveness in handling class imbalances using the SMOTE algorithm. Figure 3 can guide the selection of the most suitable classifier for the given dataset and task, considering both overall performance and specific evaluation metrics.

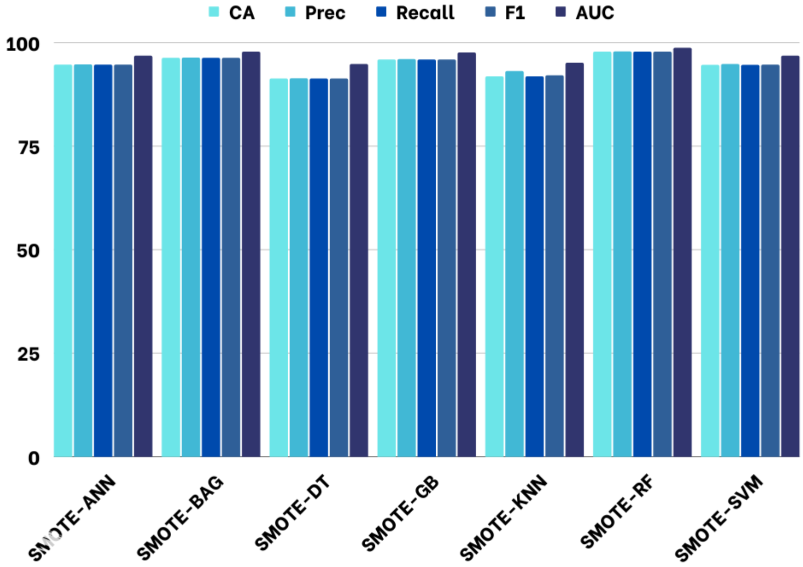

Figure 4: Classification results comparison using SMOTE

To compare the CA and AUC of models using SMOTE and without using SMOTE, we present the following Figure 4 and Figure 5.

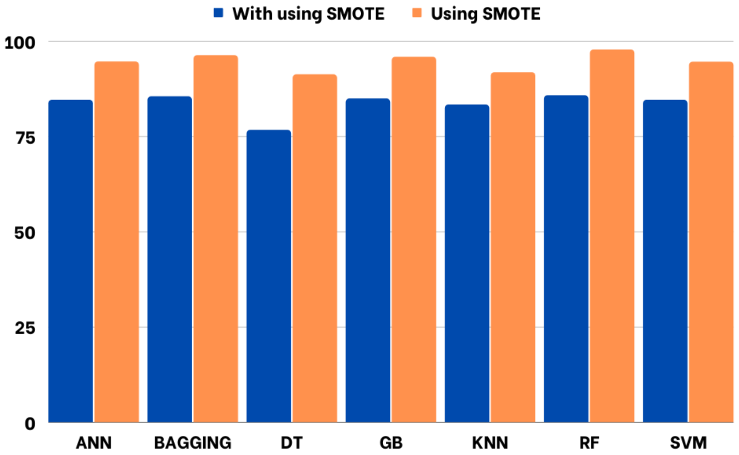

Figure 5: Comparison CA of models using SMOTE and without using SMOTE

Figure 5 shows that we observe that using SMOTE generally improves the CA of the models compared to not using SMOTE. The models consistently show higher CA values when SMOTE is applied. For example, the RF model achieves a CA of 85.73% without SMOTE, which increases to 97.75% with SMOTE.

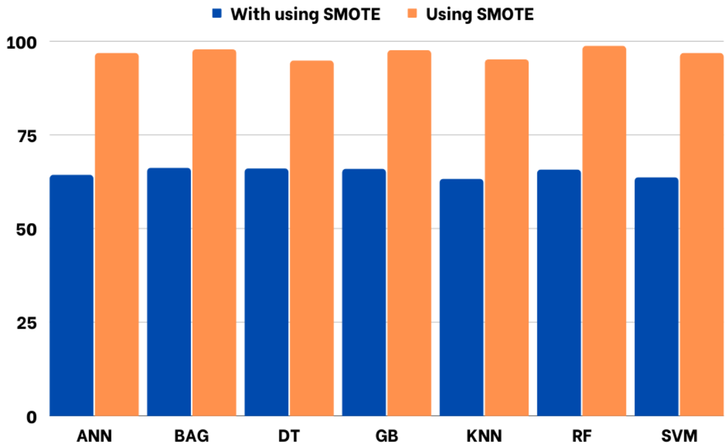

Figure 6: Comparison AUC of models using SMOTE and without using SMOTE

From Figure 6, we can see that using SMOTE also leads to higher AUC values across the models. AUC measures the overall performance of the models in distinguishing between positive and negative instances, considering various thresholds. The AUC values show significant improvements when SMOTE is utilized. For instance, the RF model achieves an AUC of 65.66% without SMOTE, which increases to 98.65% with SMOTE.

These results indicate that applying SMOTE helps to improve both the CA and AUC of the models. It effectively addresses class imbalance, allowing the models to better capture the patterns and make more accurate predictions. The substantial improvements in AUC suggest that SMOTE enhances the models' ability to rank instances correctly and distinguish between classes, leading to improved overall classification performance.

Although the experimental results have very good results for the application of machine learning to hemogram blood test samples, However, one limitation of this study is that the dataset used for testing was collected from a single hospital in a specific geographic region. This may limit the generalizability of the findings to other populations or settings. Another limitation of the study is that the collected dataset only includes six diseases from the LIS database. This may limit the applicability of the findings to a wider range of diseases. Further research is needed to evaluate the proposed approach using datasets that include a more diverse range of diseases. Additionally, the study focused solely on hemogram blood test samples and did not incorporate other types of medical data. Further research is needed to evaluate the proposed approach using larger and more diverse datasets and to explore the potential of incorporating other types of medical data.

4 Conclusion

In this study, we proposed an efficient algorithm for disease classification based on hemogram blood test samples. We applied the SMOTE algorithm to address the issue of class imbalance in the dataset and used Random Forest as the classification algorithm. Our proposed approach achieved high accuracy in classifying different diseases based on hemogram blood test samples. Specifically, the approach achieved an accuracy of up to 97.75% and an AUC of up to 98.65%.

The results of this study demonstrate the potential of machine learning in improving the accuracy and efficiency of disease diagnosis and treatment. By analyzing hemogram blood test samples, doctors can make faster and more accurate diagnoses, improving treatment outcomes for patients and reducing healthcare costs.

Future research could focus on applying the proposed approach to larger and more diverse datasets to further validate its effectiveness. In addition, the approach could be extended to other types of medical data, such as clinical data, imaging data, and genomic data, to enable more comprehensive and personalized medical services.

References

[1] G. Zini, “Artificial intelligence in hematology,” *Hematology*, vol. 10, no. 5, pp. 393–400, 2005.

[2] S. J. MacEachern and N. D. Forkert, “Machine learning for precision medicine,” *Genome*, vol. 64, no. 4, pp. 416–425, 2021.

[3] P.-H. Huynh, T.-N. Tran, and others, “Enhancing COVID-19 prediction using transfer learning from Chest X-ray images,” in *2021 8th NAFOSTED Conference on Information and Computer Science (NICS)*, IEEE, 2021, pp. 398–403.

[4] P. H. Huynh and V. H. Nguyen, “A Novel Ensemble of Support Vector Machines for Improving Medical Data Classification,” *Eng. Innov.*, vol. 4, pp. 47–66, 2023.

[5] F. K. Alsheref and W. H. Gomaa, “Blood diseases detection using classical machine learning algorithms,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, 2019.

[6] F. Akter, M. A. Hossin, G. M. Daiyan, M. M. Hossain, and others, “Classification of hematological data using data mining technique to predict diseases,” *J. Comput. Commun.*, vol. 6, no. 04, p. 76, 2018.

[7] A. Akhtar, S. Akhtar, B. Bakhtawar, A. A. Kashif, N. Aziz, and M. S. Javeid, “COVID-19 detection from CBC using machine learning techniques,” *Int. J. Technol. Innov. Manag. IJTIM*, vol. 1, no. 2, pp. 65–78, 2021.

[8] R. I. Doewes, R. Nair, and T. Sharma, “Diagnosis of COVID-19 through blood sample using ensemble genetic algorithms and machine learning classifier,” *World J. Eng.*, vol. 19, no. 2, pp. 175–182, 2022.

[9] S. Vijayarani and S. Sudha, “An efficient clustering algorithm for predicting diseases from hemogram blood test samples,” *Indian J. Sci. Technol.*, vol. 8, no. 17, p. 1, 2015.

[10] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[12] F. A. Vinisha and L. Sujihelen, “Study on Missing Values and Outlier Detection in Concurrence with Data Quality Enhancement for Efficient Data Processing,” in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, 2022, pp. 1600–1607.

[13] C. Cortes and V. Vapnik, “Support vector machine,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[14] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression T rees (Monterey, California: Wadsworth)*. Inc, 1984.

[15] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[16] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Ann. Stat.*, pp. 1189–1232, 2001.

[17] E. Fix and J. Hodges, “Discriminatory analysis-nonparametric discrimination: Small sample performance,” California Univ. Berkeley, 1952.

[18] Z. Vujović, “Classification model evaluation metrics,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 599–606, 2021.

[19] Q. Wang, T.-T. Nguyen, J. Z. Huang, and T. T. Nguyen, “An efficient random forests algorithm for high dimensional data classification,” *Adv. Data Anal. Classif.*, pp. 1–20, 2018.

[20] M. Zhu *et al.*, “Class weights random forest algorithm for processing class imbalanced medical data,” *IEEE Access*, vol. 6, pp. 4641–4652, 2018.

[21] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines,” *ACM Trans. Intell. Syst. Technol. TIST*, vol. 2, no. 3, p. 27, 2011.

[22] Y. Qi, “Random forest for bioinformatics,” *Ensemble Mach. Learn. Methods Appl.*, pp. 307–323, 2012.