Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 12 NO: 4 PAGE: 1261-1274 YEAR: 2023
DOI:10.17798/bitlisfen.1376817

Enhancing Early Detection of Blood Disorders through A Novel Hybrid Modeling Approach

Pınar KARADAYI ATAŞ [1*]

[1] *Istanbul Arel University, Faculty of Engineering and Architecture, Department of Computer Engineering, Istanbul, Türkiye*
(ORCID: 0000-0002-9429-8463)

Keywords: applied statistic, statistical analysis and application, structural and functional data, machine learning.

Abstract
Blood disorders are such conditions that impact the blood's ability to function correctly. There is a range of different symptoms depending on the type. There are several different types of blood disorders such as Leukemia, chronic myelocytic leukemia, lymphoma, myelofibrosis, polycythemia, thrombocytopenia, anemia, and leukocytosis. Some resolve completely with therapy or do not cause symptoms and do not affect overall lifespan. Some are chronic and lifelong but do not affect how an individual life. Other blood disorders, like sickle cell disease and blood cancers, can be even fatal. There needs to be a capture of hidden information in the medical data for detecting diseases in the early stages. This paper presents a novel hybrid modeling strategy that makes use of the synergy between two methods with histogram-based gradient boosting classifier tree and random subspace. It should be emphasized that the combination of these two models is being employed in this study for the first time. This novel model is presented for the assessment of blood diseases. The results show that the proposed model can predict the tumor of blood disease better than the other classifiers.

1. Introduction

Providing accessible and reliable diagnosis is a fundamental problem for global healthcare systems [1]. Early diagnosis of blood diseases is critical for both successful treatment and the avoidance of misdiagnoses, which are sadly frequent in medical practice. A misdiagnosis may result in ineffective treatment strategies, postponed care, and in extreme circumstances, serious health decline. This is particularly important when discussing blood disorders because their symptoms might be mild and easily missed or confused with other illnesses [2]. An estimated 5% of outpatients in the US alone are given the incorrect diagnosis each year. Since one in three misdiagnoses of patients with major medical illnesses results in serious patient harm, it is estimated that 20% of patients with serious medical disorders receive incorrect diagnoses at the primary care level [3].

Improving the precision and scope of diagnostic processes is critical in the field of medicine, especially for blood disorders. A multidisciplinary approach is necessary to fully comprehend the patient's condition [4]. In order to augment the information currently available, more data is gathered during a diagnostic procedure from the patient's medical history, physical examination, and various diagnostic techniques, such as clinical laboratory tests. Laboratory tests are used to guide medical care as well as to confirm, rule out, classify, or monitor illnesses. [5]. The entire potency of laboratory test results is often overestimated since clinical laboratories often publish test results as individual numerical or categorical values, and clinicians typically focus on those values that fall outside of a given reference range.

* Corresponding author: pinaratas@arel.edu.tr
Received: 16.10.2023, Accepted: 04.12.2023

The development of advanced computational techniques has transformed the study and practice of medicine, providing new opportunities for the analysis of intricate health data [6]–[10]. For many years, medical datasets have been analyzed using machine learning techniques. [11]. It provides a number of essential tools for machine learning-based intelligent data analysis. Patient monitoring and other data gathering tools are now widely used in modern hospitals to gather data, which is then shared and kept in extensive information systems. These days, machine learning technology is ideal for identifying illnesses and evaluating medical data. Clinical data analysis reveals the molecular processes that underlie diseases and the ways in which risk factors impact their progression. Medical records in hospitals or their departments contain information about correct diagnosis. Patient data with the right diagnosis are loaded into a computer program and a learning algorithm is executed in medical machine learning studies. [11]. Machine learning (ML) has the potential to greatly improve the accuracy, efficiency, and dependability of systems intended for the diagnosis of particular illnesses[12].

Laboratory blood testing is the mainstay for the clinical diagnosis of hematological disorders; however, even the most skilled hematologist may overlook trends, outliers, or correlations among the myriad blood parameters that modern laboratories are now measuring [13], [14] . In contrast, this field of medicine is particularly interesting for machine learning applications because machine learning algorithms can manage hundreds of attributes (parameters) and can identify and utilize the interactions between these many attributes.

Machine learning has demonstrated potential in a number of areas related to medical practice, such as improving differential diagnosis, helping to choose the best course of treatment, offering prognostic estimates, lowering medical errors, and increasing overall effectiveness. Its uses in hematology are steadily growing in three main domains: image interpretation, diagnostic procedures, and predictive modeling.

Predictive modeling is one application of machine learning in hematology. A noteworthy study by [15] created a model to predict 100-day mortality after allogeneic hematopoietic stem cell transplantation (HSCT) using 28,236 acute leukemia patients from the European Society of Blood and Marrow Transplantation registry. This model was validated both internally and externally, showcasing machine learning's capacity to offer insightful information for crucial medical decisions.

An interpretable boosted decision tree model that performed better than the previous benchmark for outcome prediction was used in the model's creation. [16] used similar techniques to predict acute graft-versus-host disease (GVHD). Other groups have focused on creating techniques that use imaging and gene expression data to predict treatment response [17]. Artificial intelligence decision support solutions for oncology are already available. A model developed by [18] generates a list of probable diagnoses based on age, serial chemistry, and complete blood count laboratory values. These diagnoses are then fed into a support vector machine model. The aforementioned findings suggest several possible real-world uses for artificial intelligence.

IBM Watson for Oncology ranks and suggests treatment options based on patient and illness characteristics, published literature, available clinical trials, and the expertise of top oncologists. It uses the EMR's natural language processing and machine learning algorithms to achieve this [19]. Numerous techniques have been used to apply AI to improve the efficacy, practicality, or accuracy of diagnoses. It has been demonstrated that CNN-based techniques can reliably identify multiple myeloma based only on mass spectrometry data from peripheral blood [20].

Machine learning has great potential for the field of hematology as well as the larger medical community, even though it is still in its infancy [21]. This review intends to clarify important artificial intelligence (AI) concepts for readers who are not familiar with the field, examine the various hematology applications where AI has proven useful, and talk about the new difficulties that arise when incorporating AI into clinical practice. Moreover, the aim is to offer perspectives on how these cutting-edge technologies might influence clinical outcomes and patient care in the future [22]. Machine learning has been used in some research to forecast the number of instances of a specific disease in a given region based on historical data and present conditions [23]. Others have utilized machine learning to determine the most likely sources of an outbreak based on the pathogen's genetic makeup and infection pattern [24], [25]. Others have utilized machine learning to estimate an individual's risk of developing an infectious

disease based on their persona[15]l attributes and activities [26].

The use of machine learning in infectious illness prediction is a promising area of research with potential implications in public health, epidemiology, and clinical practice. However, there are substantial obstacles and constraints to utilizing machine learning in this context, such as the requirement for high-quality data, the complexity of the underlying phenomenon, and the risk of bias and overfitting [27].

Das and Tsanas et al. [28], [29] developed novel artificial intelligence-based methodologies for analyzing Parkinson's disease patients. Little et al. [30] proposed using speech signal information to distinguish Parkinson's disease. They discriminated between 23 Parkinson's disease patients and 8 healthy people. SVM is used to define both Parkinson's disease and healthy people. The proposed technique was found to be 91.4% accurate. Another survey [28] chose 132 elements based on dysphonic discourse indicators. Specular selection (FS) calculations such as LASSO, Relief, MRMR, and LLBFS[29].

A crucial component use in blood disease detection research, image datasets have certain drawbacks, including reliance on high-quality imaging and interpretive variability [31]–[36]. For image processing, these techniques also demand a substantial amount of processing power. On the other hand, our work offers clear benefits as it makes use of numerical data from laboratory tests. Subjectivity in image analysis is reduced when dealing with numerical data because they are more standardized and structured. They offer measurable metrics, which are essential for reliable diagnosis. This method also makes it simpler to integrate patient data with other sources, which supports diagnostic models that are more thorough. Additionally, models based on numerical data need less processing power, which makes them more accessible and useful in a range of clinical contexts. In hematology, this efficiency is essential for prompt and precise diagnosis, which results in more efficient treatment planning.

A method for monitoring blood pressure (BP), which is essential for identifying and averting health problems like hypertension and cardiovascular disorders, is presented in a study in the field [37] . With the use of CNN-LSTM and Photoplethysmography (PPG) signals, the study successfully divides blood pressure (BP) into three groups: normotension, prehypertension, and hypertension. It is noteworthy that it distinguishes between normotension and hypertension with a noteworthy accuracy of 66.76%, highlighting the potential of sophisticated techniques in continuous blood pressure monitoring.

Five machine learning models (RF, NB, LogR, SVM, and AdaBoost) were developed by Kim, T., et al. using a clinical database to predict persistent immune thrombocytopenia in pediatric ITP patients. The study comprised 969 juvenile patients with ITP, of which 332 had verified acute ITP and 253 had chronic ITP. In order to predict chronic ITP, 10-fold cross-validation was carried out using clinical (age, gender, race, ethnicity, presence of primary ITP) and laboratory variables (baseline platelet count, leukocyte count, lymphocyte count, eosinophil count, mean platelet volume, anti-nuclear antibody, immature platelet fraction, direct antiglobulin test, and immunoglobin levels). When it came to predicting chronic ITP, the 100-tree random forest model performed better than any other model (AUC: 0.795, accuracy: 0.737, precision: 0.738, F1-score: 0.671, and recall: 0.737). Naïve Bayes was the second-best performing model (AUC: 0.792, accuracy, 0.698, precision: 0.737, F1-score: 0.671, and recall: 0.698) [38]

Seven machine learning models were developed to predict hospital-acquired thrombocytopenia (HAT) post-surgery in the study by Cheng et al. These models included Gradient Boosting (GB), Random Forest (RF), Logistic Regression (LogR), XGBoost, Multilayer Perceptron, Support Vector Machine (SVM), and k-nearest Neighbors (k-NN). Adult ICU patients who had undergone surgery were enrolled in the study, with training and assessment divided 70-30. In roughly 13.1% of cases, thrombocytopenia occurred. The RF and GB models fared the best in internal validation, exhibiting high levels of specificity (79.1% and 73.7%, respectively) and sensitivity (79.3% and 73.6%, respectively), with no discernible differences between them. RF and GB had AUCs of 0.834 and 0.828, respectively [39]. In order to predict 30-day mortality in ITP patients with cerebral bleeding, Zhang, X.H. et al. [40] created ten machine learning algorithms (ML): SVM, k-NN, LogR, linear discriminant analysis, decision tree, RF, GB decision tree, AdaBoost, XGBoost, and light gradient boosting machine. In the training cohort, they carried out a 10-fold cross-validation, and they externally verified across 11 different centers. During internal validation, the SVM model performed better in predicting 30-day mortality (AUC: 0.879, F-1 score: 0.748, sensitivity: 0.600).

A crucial component of medical diagnostics, early identification of blood disorders has a major influence on patient outcomes [41]. Blood disorders can have a significant impact on general health and quality of life [42]. They include a broad range of conditions from anemia to leukemia. Early detection of these conditions is essential for the possibility of early intervention, which could reduce the severity of the illness and enhance the prognosis [43].

Unfortunately, there are a number of drawbacks to the current blood disorder diagnostic techniques [44]. Even though they work well, traditional blood tests frequently miss subtle symptoms of early-stage disorders or fail to present a complete picture of the patient's health. In addition, a number of blood disorders share symptoms with other illnesses, which could result in a delayed or incorrect diagnosis.

The gaps in the diagnostic procedures used today emphasize the need for more sophisticated and accurate detection techniques [45]–[47]. By facilitating prompt and focused treatments, an improved strategy for early detection lowers the risk of complications while also increasing the accuracy of diagnoses. Closing these gaps and improving patient care in the field of blood disorders requires the development and application of novel diagnostic tools. Through the introduction of a novel hybrid modeling approach that blends cutting-edge analytics with conventional diagnostic techniques, our research seeks to address these issues and provide a more effective and efficient pathway for the early detection of blood disorders.

In this study, prior research on blood condition was assessed for early detection critically, noting important constraints and the results attained in these investigations. Gaps in present diagnostic techniques are identified by this assessment, especially with regard to the detection of subtle and early-stage signs of blood diseases. By merging random subspace techniques with gradient boosting based on histograms, our research advances this subject through the introduction of a novel hybrid modeling methodology. By addressing the inadequacies of conventional diagnostic approaches, this novel methodology seeks to improve the accuracy and reliability of blood disease diagnosis. The theory was that specific hematological diseases like leukocytosis, anemia, and thrombocytopenia found in the values of blood test results would be enough for the novel hybrid predictive model to suggest a plausible diagnosis if it were trained on a large enough dataset of medical cases that included clinical laboratory blood tests. Two separate methodologies, the Random Subspace Ensemble method and the Histogram-based Gradient Boosting Classification Tree (HIST-GBCT) algorithm, are combined to create a novel methodology. To the best of knowledge, this combination has not before been investigated in the literature.

2. Material and Method

2.1. Dataset

The 4000 samples in the dataset are categorized into three distinct groups: 1232 samples relate to pediatric hematology cases, 1451 samples are from adult hematology patients, and the remaining 1232 samples are linked to different tumor types. Sensitive personal data, including names and IDs, has been removed from the dataset in order to respect privacy standards. A salient characteristic of our dataset is the 'Clinic Number,' which functions as the classification target label. The three groups are identified by this label: 80 stands for adult hematology cases, 95 for pediatric hematology cases, and 59 for tumor cases. The National Heart, Lung, and Blood Institute's (2016) guidelines were carefully followed in the selection of the dataset's features to ensure their applicability and coherence within the framework of oncology and hematological research. The features of the dataset are listed depending on (National Heart Lung Blood Institute site, 2016). The content of the dataset's columns is presented in Table 1.

Table 1. Data set content

Name	Definition
WBC	White blood cells
RBC	RED blood cells
Cupper	Cu
hgb	Hemoglobin
HCT	Hematocrit
MCV	Macrocytic Anemia:
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
PLT	Platelet Count
RDW-SD	Red blood cell distribution width
RDW-CV	Red blood cell distribution width
PDW	platelet distribution width
MPV	Mean platelet volume
RDW-SD	Red blood cell distribution width
P-LCR	Platelet larger cell ratio
PCT	Procalcitonin
NEUT	Neutropenia

LYMPH	Lymphocytes	OUTPUT	Thrombocytopenia = 1
MONO	Mononucleosis		Normal = 2
EO	Eosinophil granulocyte		Leukocytosis = 3
BASO	Basophil granulocyte		
IG	Intravenous immunoglobulin		

2.2. Model background

Random Subspace Ensemble (RSE) is a machine learning technique used to improve predictive modeling by training several models on randomly selected subsets of the input characteristics. Each model in the ensemble focuses on a different subset of the available features, encouraging model variety. RSE tries to increase the overall model's robustness and generalization performance by merging predictions from several diverse models, making it particularly beneficial for high-dimensional datasets or when dealing with the curse of dimensionality. Ho et al. presented a classic integrated technique called random subspace in 1998 [52]. To generate the training subset, the algorithm is comparable to the bagging algorithm and is randomly selected by the original training set [52], [53]. The algorithm of the Random subspace is represented in Alg. 1.

Algorithm: Random Subspace Method

Input:

Training dataset: (*X_train*, *y_train*)
Number of base models: *n_estimators*
Maximum number of features to consider for each base model: max_features

Output:

List of base models: estimators
Initialize an empty list estimator to store the base models.
For *i* in range(*n_estimators*):

a. Randomly select a subset of features from the available features with a size not exceeding max_features. Let's call this subset selected_features.
b. Create a base model (e.g., decision tree classifier) and train it using the training data (*X_train*[:, *selected_features*], *y_train*).
c. Append the trained base model and the selected_features to the estimators list.
d. Return the list of estimators.

The Histogram-based Gradient Boosting Classification Tree (HIST-GBCT) is a supervised classification machine learning algorithm. It creates an ensemble of decision trees by enhancing their forecast accuracy iteratively. The use of histograms to efficiently represent and manipulate data during the training phase distinguishes HIST-GBCT. HIST-GBCT organizes data into bins or buckets rather than individual data points, reducing computing complexity and allowing for faster training. This method is very useful when working with huge datasets or high-dimensional feature spaces. The algorithm gradually refines its judgment bounds, resulting in a sophisticated classification model capable of successfully classifying new, previously unseen data points [54].

The outline of the mathematical formulation of Histogram-based Gradient Boosting:

Initialize the ensemble model as an empty function: $f_0(x) = 0$. For each boosting round m from 1 to M:

Step 1: Compute Negative Gradient

For Calculating the negative gradient of the loss function with respect to the current model's predictions

$$g_m(x_i) = \frac{dL\left(y_i, F_{m-1}(x_i)\right)}{dF_{m-1}(x_i)}, \text{for i} = 1,2,3,\dots,\text{N.} \quad (\mathbf{2.1})$$

Step 2: Fit a Base Learner

Fit a base learner to the negative gradients $g_m(x_i) = 0$ and discover the best structure for the tree (e.g., split points and leaf values). This is accomplished by minimizing a loss function, such as squared error in regression or deviation in classification.

$$Tree_m = argmin_{tree} \sum_{i=1}^{N} L\left(y_i, F_{m-1}(x_i) + tree(x_i)\right). (\mathbf{2.2})$$

Step 3: Update the Model

$$F_m(x) = y_i, F_{m-1}(x) + tree(x_i) + \eta. tree_m(x_i) \ (\mathbf{2.3})$$

Add the new tree to the ensemble model, weighted by a learning rate:

Step 4: Termination Check

Steps 1-3 must be repeated for a set number of boosting rounds M or until a convergence condition is fulfilled.

2.3. Proposed model

Figure 1 depicts the classification process with the whole flow of operations through the various stages. Two major operations were carried out in the preprocessing step to prepare the dataset for future analysis that represented under phase 1.

First, categorization was employed, which included converting non-numeric categorical information into a numerical format that can be used in machine learning models. This phase guarantees that the algorithms can work effectively with categorical data, contributing to the dataset's overall quality.

Accordingly, feature normalization was used. Normalization of features is an important step in data preprocessing. Scaling the numerical properties to a common range or distribution is involved. This normalization procedure guarantees that all features have a comparable influence on the modeling process, avoiding particular features from dominating the others due to scale disparities.

Furthermore, the dataset has been divided into two parts: the training set and the test set. The dataset was randomly divided into two subsets, with 70% going to the training set and 30% going to the test set. Because of the random split, both sets are representative of the whole dataset, allowing for robust model training and evaluation. This section is necessary for model training and evaluation. The training set is used to train machine learning models, while the test set is used to evaluate their performance and generalizability.

Under phase 2; the efficacy of a random forest-based method for selecting relevant features, thus enhancing data input quality, is being investigated. The Histogram-based Gradient Boosting Classification Tree (HIST-GBCT) technique is employed in an ensemble of random subspaces. This method allows us to efficiently classify the selected features, which contributes to better analysis accuracy.

In phase 3; a comprehensive set of experiments is carried out to demonstrate the efficacy of the proposed methods. These studies are designed to empirically show the benefits and efficacy of the chosen strategies.

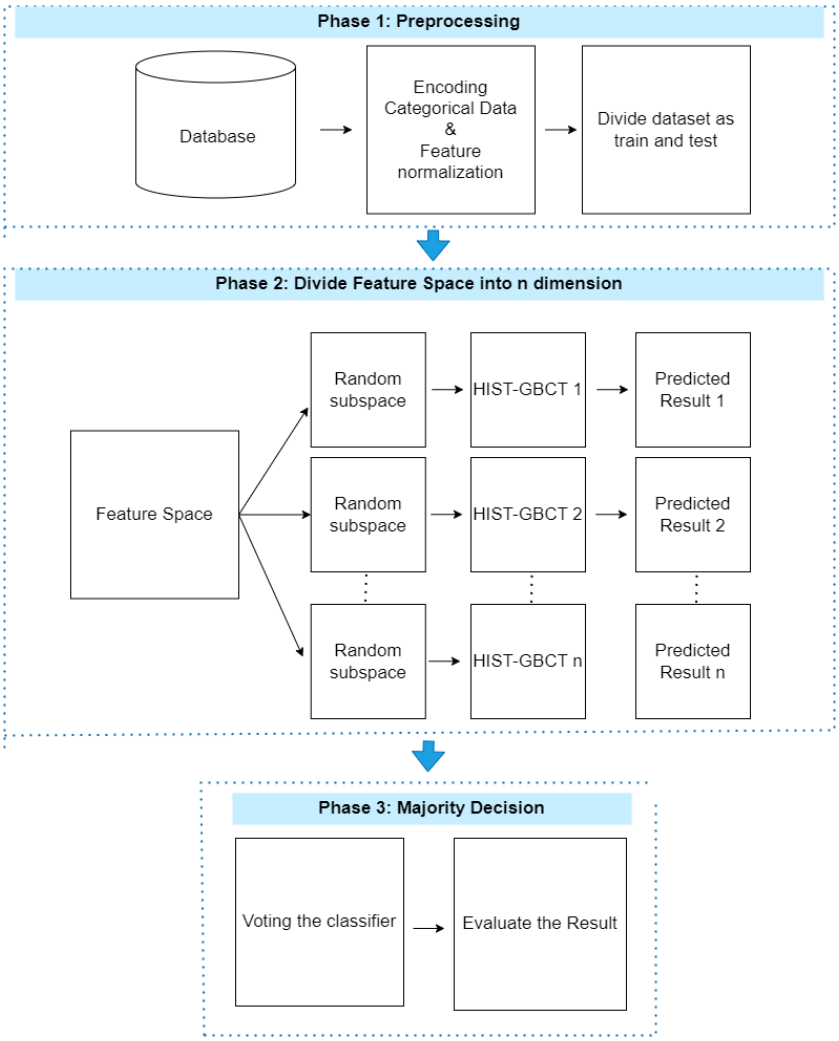

Figure 1. Illustration of the proposed workflow.

3. Results and Discussion

In this study, the focus is placed on the classification of patients with thrombocytopenia, leukocytosis, and individuals with normal hematological profiles by utilizing the proposed model. Following that, comparison analysis is performed by comparing the findings to those produced by established approaches such as logistic regression, Bernoulli Naive Bayes, SVM-linear, and k-Nearest Neighbors (kNN). Table 2 demonstrates unequivocally that the proposed model has greater predictive performance [55].

Table 2. Comparative Performance Analysis

Model Name	Accuracy	Precision	Recall	F1 score	AUC
Logistic Regression	0.842	0.841	0.802	0.799	0.694
Bernoulli Naive Bayes	0.641	0.638	0.575	0.516	0.800
SVM-Linear	0.635	0.631	0.623	0.490	0.771
KNN	0.820	0.819	0.741	0.755	0.721
Proposed Model	**0.896**	**0.894**	**0.864**	**0.862**	**0.967**

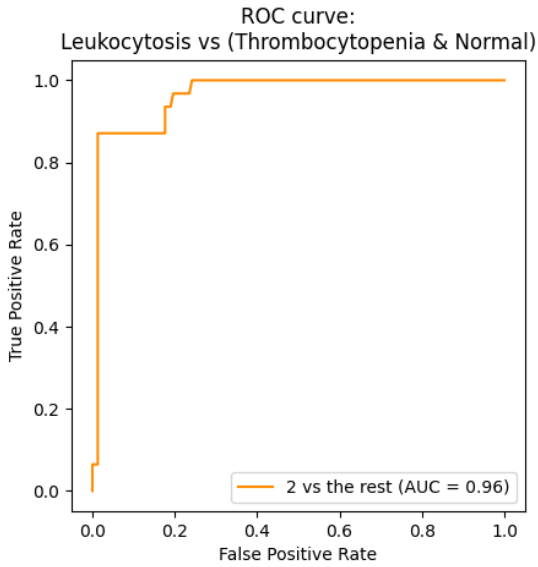

Figure 1. ROC curve for Leukocytosis vs (Thrombocytopenia and Normal).

Figure 1 shows the ROC curve displaying the classification of blood illnesses into two categories, those are 'Leukocytosis' (class_id=2) and 'Thrombocytopenia and Normal' (all other classes). Figure 2 depicts the ROC curve for classifying blood ailments into two groups: 'Thrombocytopenia (class_id=2) and 'Leukocytosis and Normal' (all other classes).

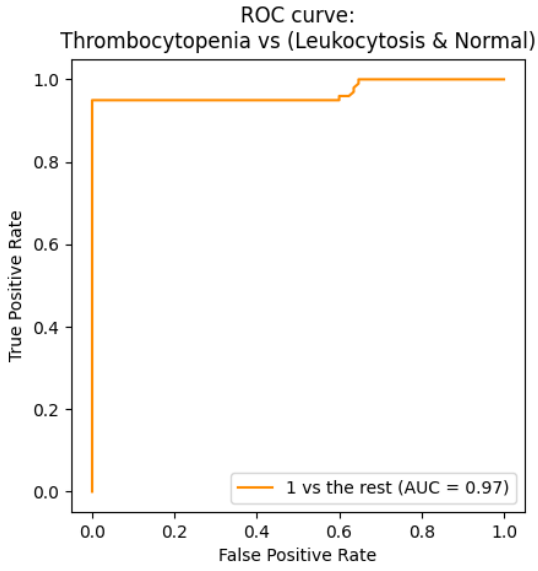

Figure 2. ROC curve for Thrombocytopenia vs (Leukocytosis and Normal).

The suggested technique regularly achieves superior ROC curve findings, suggesting its strong performance in accurately differentiating between Leukocytosis and 'Thrombocytopenia and Normal.' These excellent ROC curves highlight the method's usefulness in achieving high sensitivity and specificity, which are critical for accurate disease categorization.

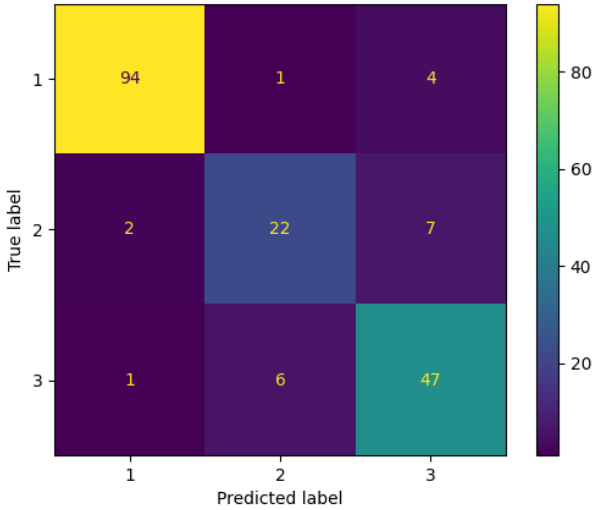

Figure 3. Confusion matrix results

Figure 3 illustrates the results of the confusion matrix. The high-performance confusion matrix demonstrates the several advantages of the technique. It emphasizes not just the model's ability to reliably identify examples, but also its ability to decrease misclassifications. As a result, diagnostic accuracy improves, false positives are minimized, and overall clinical decision support improves, ultimately contributing to more effective disease management and patient care.

Table 3 shows a comparative study of the performance of our suggested model on different sized datasets. Important performance indicators like training time, prediction time, accuracy, and memory utilization are highlighted in the table. Three different dataset sizes—small (800 rows), medium (1000 rows), and big (1400 rows)—are assessed using these metrics. The outcomes show how the model scales with increasing data quantity in terms of accuracy and efficiency, offering important information about how well-suited it is for various clinical circumstances and dataset complexity. The value of second is averaged.

Table 3: Performance Analysis of the Proposed Model for Varying Dataset Sizes

Dataset Size	Model	Training Time	Prediction Time	Accuracy	Memory Usage (Training)	Memory Usage (Prediction)
Small (1100 rows)	Proposed Model	10 seconds	1 second	83%	204 MB	50 MB
Medium (2200 rows)	Proposed Model	12 seconds	2 seconds	85%	411 MB	96 MB
Large (4000 rows)	Proposed Model	15 seconds	4 seconds	89%	622 MB	231 MB

In this paper, a novel hybrid modeling strategy was presented that takes advantage of the synergistic advantages of two different methods: the histogram-based gradient boosting classifier tree and the random subspace approach. The major goal was to develop a prediction model optimized for the assessment of blood disorders, with a particular emphasis on tumor detection.

Table 4 offers a thorough summary of the most important hematology-related studies that use machine learning methods. It contains each study's title, author names, scope, particular machine learning techniques employed, and key conclusions or contributions from each study. The table also highlights the progress and difficulties in applying machine learning to different hematological aspects by contrasting and comparing the methods and results of these studies. This compilation is a useful tool for learning about the state of machine learning applications in this important field of medical research as well as their future potential.

Table 4: Compilation and comparsion of Machine Learning Studies in Hematology

References	**Objective**	**Dataset**	**Method**	**Performance**
[48]	Predict acute lymphoblastic leukemia (ALL)	336 diagnosed children with ALL	Random forest algorithm	Accuracy: 0.829 AUC: 0.902
[49]	Predict chronic myeloid leukemia (CML)	Complete blood count records of 1623 people with CML	XGBoost and LASSO	AUC range: 0.87-0.96
[50]	Detection of leukemia and its types	220 blood smear images from healthy individuals and patients with leukemia	support vector machine	Accuracy: 0.80
[51]	Leukemia image segmentation	The Acute Lymphoblastic Leukemia Image Database	HSCRKMb/particle swarm optimization/K-means	Accuracy: 0.80
[17]	Prediction of complete remission of acute myeloid leukemia	473 bone marrow samples	K-nearest neighbor, support vector machine, and hill climbing	AUC :0.84
Proposed Model	**Prediction the tumor of blood disease**	**The 4000 samples for hematology**	**The Histogram-based Gradient Boosting Classification Tree with Random Subspace Method**	**Accuracy :0.896 Precision :0.894 Recall :0.864 F1 score:0.862 AUC :0.967**

The findings show that suggested model outperforms other classifiers in predicting blood disease tumors. This accomplishment highlights the study's approach potential therapeutic relevance and utility in assisting medical practitioners in the early detection of these crucial illnesses.

This study's methodology is unique in that it combines the Random Subspace Method with the Histogram-based Gradient Boosting Classification Tree. The goal of this special pairing, which hasn't been discussed in the literature before, is to maximize the advantages of both approaches. Although the Gradient Boosting based on Histogram provides a strong instrument for managing intricate data structures, the Random Subspace Method improves the model's capacity to generalize and function in various diagnostic situations. This technology offers a more sophisticated and useful tool for early blood problem diagnosis, marking a substantial shift from conventional diagnostic techniques.

It is worth noting that the encouraging results of this study pave the path for additional research and therapeutic applications. The accuracy and effectiveness of the model suggest that it could be a useful tool in the field of healthcare, particularly for the early diagnosis of blood diseases. However, it is critical to recognize the study's limitations, such as the size of the dataset and the necessity for real-world validation.

4. Conclusion

Blood disorders are a broad category of medical illnesses that can have a significant impact on the correct functioning of the circulatory system. This broad range of illnesses produces a variety of symptoms that can range from minor to severe, depending on the exact type and unique patient features. Among the many blood illnesses include leukemia, chronic myelocytic leukemia, lymphoma, myelofibrosis, polycythemia, thrombocytopenia, anemia, and leukocytosis. Understanding these illnesses is critical because they have a wide range of consequences, from those that can be adequately managed to those that can be fatal.

Early detection of blood problems is one of the most difficult challenges in the field of blood disorders. An accurate diagnosis is critical for commencing appropriate therapeutic interventions, which can have a major impact on patient outcomes. However, due to the intricacy of the underlying illness processes, detecting blood problems in their early stages can be extremely difficult.

In conclusion, the findings of the study show the utility of hybrid modeling strategies in the context of blood condition assessment, particularly tumor prediction. These findings lay the groundwork for future research efforts aimed at improving early detection, treatment, and overall outcomes for those affected by blood diseases.

Acknowledgement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for profit sectors.

Author's Contributions

Ethics *(mandatory)*

There are no ethical issues after the publication of this manuscript.

References

[1] H. Singh, A. N. D. Meyer, and E. J. Thomas, 'The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations', *BMJ Qual. Saf.*, vol. 23, no. 9, pp. 727–731, Sep. 2014, doi: 10.1136/bmjqs-2013-002627.

[2] T. M. Ghazal, A. U. Rehman, M. Saleem, M. Ahmad, S. Ahmad, and F. Mehmood, 'Intelligent Model to Predict Early Liver Disease using Machine Learning Technique', presented at the 2022 International Conference on Business Analytics for Technology and Security (ICBATS), IEEE, 2022, pp. 1–5.

[3] M. L. Graber, 'The incidence of diagnostic error in medicine', *BMJ Qual. Saf.*, vol. 22, no. Suppl 2, pp. ii21–ii27, Oct. 2013, doi: 10.1136/bmjqs-2012-001615.

[4] J. J. Deeks, P. M. Bossuyt, M. M. Leeflang, and Y. Takwoingi, *Cochrane handbook for systematic reviews of diagnostic test accuracy*. John Wiley & Sons, 2023.

[5] T. Badrick, 'Biological variation: Understanding why it is so important?', *Pract. Lab. Med.*, vol. 23, p. e00199, Jan. 2021, doi: 10.1016/j.plabm.2020.e00199.

[6] A. Alanazi, 'Using machine learning for healthcare challenges and opportunities', *Inform. Med. Unlocked*, vol. 30, p. 100924, 2022.

[7] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, and S. A. Tsaftaris, 'Causal machine learning for healthcare and precision medicine', *R. Soc. Open Sci.*, vol. 9, no. 8, p. 220638, 2022.

[8] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, 'Significance of machine learning in healthcare: Features, pillars and applications', *Int. J. Intell. Netw.*, vol. 3, pp. 58–73, 2022.

[9] S. Aminizadeh *et al.*, 'The applications of machine learning techniques in medical data processing based on distributed computing and the Internet of Things', *Comput. Methods Programs Biomed.*, p. 107745, 2023.

[10] A. Y. Gill, A. Saeed, S. Rasool, A. Husnain, and H. K. Hussain, 'Revolutionizing Healthcare: How Machine Learning is Transforming Patient Diagnoses-a Comprehensive Review of AI's Impact on Medical Diagnosis', *J. World Sci.*, vol. 2, no. 10, pp. 1638–1652, 2023.

[11] M. Shehab *et al.*, 'Machine learning in medical applications: A review of state-of-the-art methods', *Comput. Biol. Med.*, vol. 145, p. 105458, Jun. 2022, doi: 10.1016/j.compbiomed.2022.105458.

[12] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, 'Multiple disease prediction using Machine learning algorithms', *Mater. Today Proc.*, vol. 80, pp. 3682–3685, 2023.

[13] R. J. Means Jr *et al.*, *Wintrobe's clinical hematology*. Lippincott Williams & Wilkins, 2023.

[14] M. Auerbach, 'Optimizing diagnosis and treatment of iron deficiency and iron deficiency anemia in women and girls of reproductive age: clinical opinion', *Int. J. Gynecol. Obstet.*, vol. 162, pp. 68–77, 2023.

[15] R. Shouval *et al.*, 'Validation of the acute leukemia-EBMT score for prediction of mortality following allogeneic stem cell transplantation in a multi-center GITMO cohort', *Am. J. Hematol.*, vol. 92, no. 5, pp. 429–434, May 2017, doi: 10.1002/ajh.24677.

[16] Y. Arai *et al.*, 'Using a machine learning algorithm to predict acute graft-versus-host disease following allogeneic transplantation', *Blood Adv.*, vol. 3, no. 22, pp. 3626–3634, Nov. 2019, doi: 10.1182/bloodadvances.2019000934.

[17] O. Gal, N. Auslander, Y. Fan, and D. Meerzaman, 'Predicting Complete Remission of Acute Myeloid Leukemia: Machine Learning Applied to Gene Expression', *Cancer Inform.*, vol. 18, p. 117693511983554, Jan. 2019, doi: 10.1177/1176935119835544.

[18] G. Gunčar *et al.*, 'An application of machine learning to haematological diagnosis', *Sci. Rep.*, vol. 8, no. 1, p. 411, Jan. 2018, doi: 10.1038/s41598-017-18564-8.

[19] J. L. Malin, 'Envisioning Watson As a Rapid-Learning System for Oncology', *J. Oncol. Pract.*, vol. 9, no. 3, pp. 155–157, May 2013, doi: 10.1200/JOP.2013.001021.

[20] M. Deulofeu *et al.*, 'Rapid discrimination of multiple myeloma patients by artificial neural networks coupled with mass spectrometry of peripheral blood plasma', *Sci. Rep.*, vol. 9, no. 1, p. 7975, May 2019, doi: 10.1038/s41598-019-44215-1.

[21] C. J. Haug and J. M. Drazen, 'Artificial intelligence and machine learning in clinical medicine, 2023', *N. Engl. J. Med.*, vol. 388, no. 13, pp. 1201–1208, 2023.

[22] S. A. Alowais *et al.*, 'Revolutionizing healthcare: the role of artificial intelligence in clinical practice', *BMC Med. Educ.*, vol. 23, no. 1, p. 689, 2023.

[23] S. Palaniappan, R. V, B. David, and P. N. S, 'Prediction of Epidemic Disease Dynamics on the Infection Risk Using Machine Learning Algorithms', *SN Comput. Sci.*, vol. 3, no. 1, p. 47, Jan. 2022, doi: 10.1007/s42979-021-00902-3.

[24] S. Roy, P. Biswas, and P. Ghosh, 'Spatiotemporal tracing of pandemic spread from infection data', *Sci. Rep.*, vol. 11, no. 1, p. 17689, Sep. 2021, doi: 10.1038/s41598-021-97207-5.

[25] R. B. Ghannam and S. M. Techtmann, 'Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring', *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 1092–1107, 2021, doi: 10.1016/j.csbj.2021.01.028.

[26] S. Yadav, M. K. Singh, and S. Pal, 'Artificial Intelligence Model for Parkinson Disease Detection Using Machine Learning Algorithms', Biomed. Mater. Devices, Mar. 2023, doi: 10.1007/s44174-023-00068-x.

[27] J. A. Roth, M. Battegay, F. Juchler, J. E. Vogt, and A. F. Widmer, 'Introduction to Machine Learning in Digital Healthcare Epidemiology', Infect. Control Hosp. Epidemiol., vol. 39, no. 12, pp. 1457–1462, Dec. 2018, doi: 10.1017/ice.2018.265.

[28] R. Das, 'A comparison of multiple classification methods for diagnosis of Parkinson disease', Expert Syst. Appl., vol. 37, no. 2, pp. 1568–1572, Mar. 2010, doi: 10.1016/j.eswa.2009.06.040.

[29] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, 'Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity', J. R. Soc. Interface, vol. 8, no. 59, pp. 842–855, Jun. 2011, doi: 10.1098/rsif.2010.0456.

[30] M. A. Little and L. O. Ramig, 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', Nat. Preced., 2008.

[31] M. K. Gourisaria, S. Das, R. Sharma, S. S. Rautaray, and M. Pandey, 'A deep learning model for malaria disease detection and analysis using deep convolutional neural networks', Int. J. Emerg. Technol., vol. 11, no. 2, pp. 699–704, 2020.

[32] N. M. Deshpande, S. Gite, and R. Aluvalu, 'A review of microscopic analysis of blood cells for disease detection with AI perspective', PeerJ Comput. Sci., vol. 7, p. e460, 2021.

[33] D. N. Patil and U. P. Khot, 'Image processing based abnormal blood cells detection', Int. J. Tech. Res. Appl., vol. 31, pp. 37–43, 2015.

[34] R. Sigit, M. M. Bachtiar, and M. I. Fikri, 'Identification of leukemia diseases based on microscopic human blood cells using image processing', presented at the 2018 International Conference on Applied Engineering (ICAE), IEEE, 2018, pp. 1–5.

[35] P. K. Das, B. Nayak, and S. Meher, 'A lightweight deep learning system for automatic detection of blood cancer', Measurement, vol. 191, p. 110762, 2022.

[36] D. O. Oyewola, E. G. Dada, S. Misra, and R. Damaševičius, 'A novel data augmentation convolutional neural network for detecting malaria parasite in blood smear images', Appl. Artif. Intell., vol. 36, no. 1, p. 2033473, 2022.

[37] K. Gupta, N. Jiwani, and N. Afreen, 'Blood pressure detection using CNN-LSTM model', presented at the 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 2022, pp. 262–366.

[38] T. O. Kim et al., ‘Predicting Chronic Immune Thrombocytopenia in Pediatric Patients at Disease Presentation: Leveraging Clinical and Laboratory Characteristics Via Machine Learning Models’, Blood, vol. 138, p. 1023, 2021.

[39] Y. Cheng et al., ‘Using Machine Learning Algorithms to Predict Hospital Acquired Thrombocytopenia after Operation in the Intensive Care Unit: A Retrospective Cohort Study’, Diagnostics, vol. 11, no. 9, p. 1614, 2021.

[40] X.-H. Zhang et al., ‘P1652: Machine-Learning-Based Mortality Predıction of Ich In Adults With Itp: A Natıonwıde Representatıve Multicentre Study’, HemaSphere, vol. 6, no. Suppl, 2022.

[41] Y. Zhou et al., ‘Severe anemia, severe leukopenia, and severe thrombocytopenia of amphotericin B deoxycholate-based induction therapy in patients with HIV-associated talaromycosis: a subgroup analysis of a prospective multicenter cohort study’, BMC Infect. Dis., vol. 23, no. 1, p. 707, 2023.

[42] A. T. Johnsen, D. Tholstrup, M. A. Petersen, L. Pedersen, and M. Groenvold, ‘Health related quality of life in a nationally representative sample of haematological patients’, Eur. J. Haematol., vol. 83, no. 2, pp. 139–148, 2009.

[43] U. Jäger et al., ‘Diagnosis and treatment of autoimmune hemolytic anemia in adults: Recommendations from the First International Consensus Meeting’, Blood Rev., vol. 41, p. 100648, 2020.

[44] E. Franco, K. A. Karkoska, and P. T. McGann, ‘Inherited disorders of hemoglobin: A review of old and new diagnostic methods’, Blood Cells. Mol. Dis., p. 102758, 2023.

[45] E. Grudzińska and M. Modrzejewska, ‘Modern diagnostic techniques for the assessment of ocular blood flow in myopia: current state of knowledge’, J. Ophthalmol., vol. 2018, 2018.

[46] I. Voinsky, O. Y. Fridland, A. Aran, R. E. Frye, and D. Gurwitz, ‘Machine learning-based blood RNA signature for diagnosis of autism spectrum disorder’, Int. J. Mol. Sci., vol. 24, no. 3, p. 2082, 2023.

[47] S. Abd El-Ghany, M. Elmogy, and A. A. El-Aziz, ‘Computer-Aided Diagnosis System for Blood Diseases Using EfficientNet-B3 Based on a Dynamic Learning Algorithm’, Diagnostics, vol. 13, no. 3, p. 404, 2023.

[48] L. Pan et al., ‘Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia’, Sci. Rep., vol. 7, no. 1, p. 7402, 2017.

[49] R. G. Hauser et al., ‘A machine learning model to successfully predict future diagnosis of chronic myelogenous leukemia with retrospective electronic health records data’, Am. J. Clin. Pathol., vol. 156, no. 6, pp. 1142–1148, 2021.

[50] P. Jagadev and D. H. G. Virani, "Detection of Leukemia and its Types using Image Processing and Machine Learning", 2017.

[51] H. Inbarani H., A. T. Azar, and J. G, ‘Leukemia Image Segmentation Using a Hybrid Histogram-Based Soft Covering Rough K-Means Clustering Algorithm’, Electronics, vol. 9, no. 1, p. 188, Jan. 2020, doi: 10.3390/electronics9010188.

[52] S. Kotsiantis, ‘Combining bagging, boosting, rotation forest and random subspace methods’, Artif. Intell. Rev., vol. 35, no. 3, pp. 223–240, Mar. 2011, doi: 10.1007/s10462-010-9192-8.

[53] J. Mielniczuk and P. Teisseyre, ‘Using random subspace method for prediction and variable importance assessment in linear regression’, Comput. Stat. Data Anal., vol. 71, pp. 725–742, Mar. 2014, doi: 10.1016/j.csda.2012.09.018.

[54] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, ‘A comparative analysis of gradient boosting algorithms’, Artif. Intell. Rev., vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[55] F. Bulut, "Çok Katmanlı Algılayıcılar İle Doğru Meslek Tercihi", *Anadolu Univ. J. Sci. Technol.-Appl. Sci. Eng.*, vol. 17, no. 1, Apr. 2016, doi: 10.18038/btda.45787.