Leveraging Machine Learning for Accurate Anemia Diagnosis Using Complete Blood Count Data

Gregorius Airlangga
Information System Study Program, Atma Jaya Catholic University of Indonesia, Indonesia
Email: gregorius.airlangga@atmajaya.ac.id

Article Info

Article history:
Received Feb 12th, 2024
Revised Apr 20th, 2024
Accepted May 20th, 2024

Keyword:
Anemia Diagnosis
Complete Blood Count (CBC)
Decision Tree
Machine Learning
Medical Data Analysis

ABSTRACT

Anemia, a prevalent hematologic disorder, necessitates accurate and timely diagnosis for effective management and treatment. This study explores the application of various machine learning models to classify anemia types using complete blood count (CBC) data. We evaluated multiple models, including DecisionTreeClassifier, ExtraTreeClassifier, RandomForestClassifier, ExtraTreesClassifier, XGBoost, LightGBM, and CatBoost, to identify the most effective approach for anemia diagnosis. The dataset comprised CBC data labeled with anemia diagnoses, sourced from multiple medical facilities. Rigorous data preprocessing was performed, followed by feature selection using methods such as Variance Inflation Factor (VIF), Predictive Power Score (PPS), and feature importance from ensemble models. The models were trained and evaluated using 5-fold cross-validation, with hyperparameter tuning conducted via GridSearchCV. Results demonstrated that the DecisionTreeClassifier achieved the highest balanced accuracy score of 94.17%, outperforming more complex ensemble methods. Confusion matrices validated its robust performance, highlighting its precision and recall. The study underscores the potential of simple decision tree models in medical diagnosis tasks, particularly when datasets are well-preprocessed. These findings have significant implications for clinical practice, suggesting that machine learning can enhance diagnostic accuracy and efficiency. Future work will explore advanced techniques to further improve performance and integration into clinical workflows.

Corresponding Author:
Gregorius Airlangga,
Information System Study Program,
Atma Jaya Catholic University of Indonesia
Jakarta, Indonesia
Email: gregorius.airlangga@atmajaya.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v7i2.29869

1. INTRODUCTION

Anemia is a prevalent hematologic disorder that affects millions of people globally [1]–[3]. Characterized by a deficiency in the number or quality of red blood cells (RBCs) or hemoglobin, anemia can significantly impair oxygen transport to tissues, leading to a range of clinical symptoms from fatigue and weakness to severe organ dysfunction [4]–[6]. The diagnosis and classification of anemia types are crucial for effective treatment and management [7]–[9]. Traditionally, the diagnosis of anemia involves a complete blood count (CBC), followed by manual interpretation of results by healthcare professionals [10]. However, this method is time-consuming and prone to subjective variability [11]. With advancements in data science and machine learning, there is an increasing interest in leveraging these technologies to enhance the accuracy and efficiency of anemia diagnosis [12]. In recent years, machine learning has revolutionized various fields, including healthcare, by offering robust tools for data analysis and predictive modeling [13]. Numerous studies have demonstrated the potential of machine learning algorithms in diagnosing diseases, predicting patient outcomes, and aiding in personalized medicine [14]. For anemia diagnosis, machine learning models can analyze CBC data and identify patterns that may not be apparent to human observers, leading to more accurate and timely diagnoses [15].

The urgency of this research stems from the significant public health burden posed by anemia. According to the World Health Organization (WHO), anemia affects approximately 1.62 billion people worldwide, with pregnant women and young children being the most vulnerable groups [16]. Early and accurate diagnosis is essential to prevent complications and improve patient outcomes. Traditional diagnostic methods, while effective, are limited by their reliance on expert interpretation and the inherent delays in laboratory processing [17]. Machine learning offers a promising alternative that can complement existing diagnostic practices by providing rapid, automated, and consistent analysis of CBC data [18]. The state of the art in machine learning for medical diagnosis includes a wide range of algorithms, from classical statistical models to advanced ensemble techniques. Decision trees, random forests, gradient boosting machines, and deep learning models have all been applied to various medical datasets with considerable success [19]. In the context of anemia diagnosis, several studies have explored the use of machine learning models to classify different types of anemia based on CBC parameters [20]. These models have demonstrated high accuracy and reliability, suggesting that machine learning can be a valuable tool in clinical practice.

Despite the advancements in this field, there are still gaps that need to be addressed. One major challenge is the interpretability of machine learning models. While complex models like gradient boosting and deep learning can achieve high accuracy, their decision-making processes are often opaque, making it difficult for clinicians to trust and adopt these technologies [21]. Another gap is the generalizability of models trained on specific datasets. Medical data can vary significantly across different populations and healthcare settings, and models that perform well on one dataset may not necessarily generalize to others [22]. To address these gaps, the goal of this research is to develop and evaluate machine learning models for the diagnosis of anemia using CBC data. We aim to compare the performance of various classifiers, including DecisionTreeClassifier, ExtraTreeClassifier, RandomForestClassifier, ExtraTreesClassifier, XGBClassifier, LGBMClassifier, and CatBoostClassifier. By conducting a comprehensive comparison, we seek to identify the most effective model for this task and understand the trade-offs between different approaches. Additionally, we aim to enhance model interpretability by incorporating techniques that make model predictions more transparent to healthcare professionals.

Our contribution to the field is multifaceted. First, we provide a thorough evaluation of multiple machine learning models on a real-world dataset of CBC data labeled with anemia diagnoses. Second, we propose a framework for integrating machine learning into the diagnostic workflow, highlighting practical considerations for implementation in clinical settings. Third, we explore methods to improve the interpretability of machine learning models, making them more accessible to healthcare professionals. Finally, we perform a gap analysis to identify areas where further research is needed, paving the way for future advancements in this domain. The remaining structure of this journal article is organized as follows. Section 2 discusses the methodology, including data preprocessing, model training, and evaluation procedures. Section 3 provides the results of our experiments, comparing the performance of different classifiers and analyzing their strengths and weaknesses. In addition, we discuss the implications of our findings, including potential clinical applications and limitations of our approach. Section 4 concludes the article, summarizing our contributions and suggesting directions for future research.

2. RESEARCH METHOD

The research methods section as presented in the figure 1 outlines the systematic procedures followed to conduct this study. This section covers data collection, data preprocessing, feature selection, model selection, model training and evaluation, hyperparameter tuning, and performance metrics.

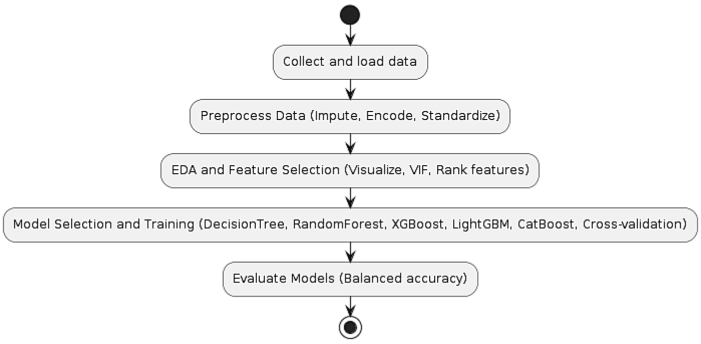
Collect and load data
Preprocess Data (Impute, Encode, Standardize)
EDA and Feature Selection (Visualize, VIF, Rank features)
Model Selection and Training (DecisionTree, RandomForest, XGBoost, LightGBM, CatBoost, Cross-validation)
Evaluate Models (Balanced accuracy)

Figure 1. Research Methodology of Anemia Diagnosis Using Machine Learning

2.1. Data Collection

The dataset used in this study comprises complete blood count (CBC) data labeled with the diagnosis of anemia type and can be colleced from [23]. This data was sourced from several medical facilities where CBCs were performed and manually diagnosed by healthcare professionals. The dataset includes a comprehensive set of attributes essential for diagnosing anemia. Hemoglobin (HGB) measures the amount of hemoglobin in the blood, which is crucial for oxygen transport. Platelets (PLT) indicate the number of platelets in the blood, which are involved in blood clotting processes. White blood cells (WBC) are counted to assess the body's immune response. Red blood cells (RBC) are counted to determine their ability to transport oxygen. The mean corpuscular volume (MCV) reflects the average volume of a single red blood cell, while the mean corpuscular hemoglobin (MCH) indicates the average amount of hemoglobin per red blood cell. The mean corpuscular hemoglobin concentration (MCHC) provides the average concentration of hemoglobin in red blood cells. Platelet distribution width (PDW) measures the variability in platelet size distribution in the blood, and procalcitonin (PCT) is a test used to diagnose sepsis or assess the risk of developing sepsis. The diagnosis attribute specifies the type of anemia based on the CBC parameters. This dataset was loaded into a pandas DataFrame for further analysis. This initial step of loading the data into a pandas DataFrame is critical for enabling subsequent data manipulation and analysis. By structuring the data in this format, we facilitate various preprocessing steps, exploratory data analysis, and the application of machine learning models. This comprehensive dataset forms the foundation of our study, providing the necessary information to develop and evaluate machine learning models for anemia diagnosis.

2.2. Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for machine learning models, ensuring that the data is clean, well-formatted, and suitable for analysis. In this study, several preprocessing steps were applied to the complete blood count (CBC) dataset to enhance the quality and utility of the data. First, the dataset was thoroughly examined for any missing values. Missing data can introduce bias and affect the performance of machine learning models. Therefore, any missing values were imputed using appropriate statistical methods. Depending on the nature and distribution of the data, either mean or median imputation was employed. Mean imputation replaces missing values with the average of the non-missing values, while median imputation uses the median value. These methods help in maintaining the integrity of the dataset by minimizing the distortion caused by missing entries. Next, the target variable, 'Diagnosis,' which indicates the type of anemia, was a categorical variable. Machine learning models typically require numerical input, so it was necessary to encode this categorical variable into numeric values. This was accomplished using the LabelEncoder from scikit-learn.

LabelEncoder converts categorical labels into integer values, thus transforming the 'Diagnosis' column into a format that can be readily processed by machine learning algorithms. Another critical step in the preprocessing pipeline was the standardization of feature variables. The CBC dataset includes various measurements such as hemoglobin levels, platelet counts, and red blood cell counts, each with different units and scales. Standardization was performed using the StandardScaler from scikit-learn, which transforms the features to have a mean of zero and a standard deviation of one. This scaling process is essential because it ensures that each feature contributes equally to the model training process, preventing features with larger scales from dominating those with smaller scales. Standardized features also help in improving the convergence speed and performance of machine learning models, particularly those that rely on distance metrics, such as support vector machines and k-nearest neighbors.

2.3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain a comprehensive understanding of the distribution and relationships among the features in the dataset. This involved generating various visualizations and analyses to uncover patterns, correlations, and potential anomalies within the data. Initially, descriptive statistics were generated using the skimpy package, providing a summary of the dataset that includes measures of central tendency, dispersion, and the overall distribution of each feature. This step offered a foundational overview, allowing for a quick assessment of the data's structure and the presence of any outliers or unusual values. To further investigate the distribution of individual features, histograms and Kernel Density Estimation (KDE) plots were created. Histograms provided a straightforward visual representation of the frequency distribution of each feature, while KDE plots offered a smoothed curve to better understand the data distribution's underlying shape. These plots were instrumental in identifying skewness, kurtosis, and other characteristics of the data. Quantile-Quantile (QQ) plots were also generated to assess the normality of the feature distributions. By comparing the quantiles of the data against a theoretical normal distribution, QQ plots helped identify deviations from normality, which is crucial for selecting appropriate statistical tests and machine learning algorithms.

Pair plots were utilized to visualize the relationships between different features. These plots provided a matrix of scatterplots for each feature pair, along with the distribution of individual features along the diagonal. Pair plots are particularly useful for identifying potential correlations and interactions between features, which can inform feature selection and engineering processes. A correlation matrix was computed using Spearman's rank correlation coefficient to quantify the strength and direction of relationships between features. This non-parametric measure was chosen for its robustness in handling non-linear relationships. A heatmap of the correlation matrix was generated to visually depict these correlations, making it easier to identify highly correlated features that might lead to multicollinearity in the machine learning models. To complement the correlation analysis, the Predictive Power Score (PPS) was calculated for each feature pair. PPS is a metric that quantifies the predictive strength of one feature on another, providing a more nuanced understanding of feature interactions. A heatmap of the PPS matrix was created, highlighting the features with the highest predictive power.

2.4. Feature Selection

Feature selection is a critical step in the machine learning pipeline, aiming to identify the most relevant features for the classification task while eliminating those that contribute little to predictive performance or introduce redundancy. In this study, a combination of statistical and model-based methods was employed to ensure a robust selection process. One of the initial methods used was the calculation of the Variance Inflation Factor (VIF) for each feature. VIF measures the extent of multicollinearity in the features, which occurs when two or more features are highly correlated and can distort the model's understanding of the data. Features with high VIF values were scrutinized and considered for removal to reduce multicollinearity, thereby enhancing the model's interpretability and stability.

In addition to VIF, the Predictive Power Score (PPS) was utilized to assess the importance of each feature in predicting the target variable. PPS is a metric that evaluates the predictive strength of one feature over another, providing a more nuanced and flexible measure of feature relevance than traditional correlation coefficients. Features with high PPS scores were deemed significant for the prediction of anemia and were prioritized in the selection process. Model-based methods were also integral to the feature selection process. Specifically, the feature importance scores derived from tree-based models like RandomForestClassifier and XGBClassifier were analyzed. These models inherently provide measures of feature importance by evaluating the contribution of each feature to the reduction of impurity in the trees. By aggregating these importance scores across all trees in the ensemble, a ranked list of features was generated, highlighting those that had the greatest impact on the model's predictive performance.

2.5. Model Selection

The process of model selection is pivotal in machine learning research, particularly when dealing with complex medical datasets such as those used for anemia diagnosis. In this study, several machine learning models were carefully selected for comparison based on their robustness, ability to handle complex datasets, and demonstrated success in similar medical diagnosis tasks. The DecisionTreeClassifier was included for its simplicity and interpretability. Decision trees work by recursively partitioning the data space and are straightforward to visualize and understand. This model's decision-making process can be easily traced, making it a valuable tool in medical applications where model transparency is critical. The decision tree algorithm can be mathematically described by the following recursive partitioning $"\{Split\}(X) = \arg\ \max\ \_\{j,t\} \vdash[\sum\_\{i = 1\}\^\{n\} \vdash(I(x\_\{ij\} \leq t) \cdot "\{Impurity\}(S\_1) + I(x\_\{ij\} > t) \cdot "\{Impurity\}(S\_2) \dashv) \dashv "$ where Split(X) is the optimal split for feature j at threshold t, I is the indicator function, S_1 and S_2 are the subsets of the data created by the split, and Impurity is a measure such as Gini impurity or entropy.

Furthermore, the ExtraTreeClassifier, an extension of the basic decision tree, was also selected. This model builds an ensemble of extremely randomized trees, which helps in reducing overfitting and improving generalization. The ExtraTreeClassifier introduces randomness in both the choice of the cut-points and the features used for splitting, thus increasing the model's robustness against noisy data. RandomForestClassifier, a widely-used ensemble method, was chosen for its ability to handle high-dimensional data and complex interactions between features. Random forests build multiple decision trees during training and output the mean prediction of the individual trees, significantly enhancing predictive accuracy and stability. This model is particularly effective in medical datasets, where the relationships between features and outcomes can be intricate and non-linear. Random forests build multiple decision trees during training and output the mean prediction of the individual trees, significantly enhancing predictive accuracy and stability. Mathematically, a random forest can be described as $\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$ where $T_b(x)$ is the prediction of the b-th decision tree, and B is the total number of trees.

Similarly, the ExtraTreesClassifier, which builds an ensemble of trees like RandomForestClassifier but with more randomness in the splitting criterion, was included. This approach helps in further reducing variance and improving the model's ability to generalize to unseen data. The XGBClassifier, part of the XGBoost framework, was selected for its efficiency and performance. XGBoost implements gradient boosting, a powerful ensemble technique that builds models sequentially, with each new model attempting to correct the errors of the previous ones. Its regularization parameters help in preventing overfitting, making it highly suitable for medical datasets. Its regularization parameters help in preventing overfitting, making it highly suitable for medical datasets. The gradient boosting process can be formulated as $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$, where $F_m(x)$ is the prediction at iteration m, γ_m is the step size, and $h_m(x)$ is the base learner (e.g., a decision tree).

LGBMClassifier, from the LightGBM framework, was chosen for its speed and scalability. LightGBM is designed to handle large datasets efficiently and supports parallel and GPU learning. Its gradient-based one-sided sampling and exclusive feature bundling techniques make it highly efficient without compromising on accuracy. Finally, the CatBoostClassifier was included due to its ability to handle categorical features natively without extensive preprocessing. CatBoost uses ordered boosting and permutation-driven feature selection, which helps in reducing overfitting and improving model interpretability.

2.6. Model Training and Evaluation

In this study, the dataset was split into training and testing sets using K-Fold cross-validation with five folds, ensuring a robust evaluation of the models. This approach allowed each data point to be included in both the training and testing sets across different folds, providing a comprehensive assessment of the models' performance. Mathematically, K-Fold cross-validation can be represented as $\text{CV}_k = \frac{1}{k}\sum_{i=1}^{k} \text{Accuracy}_i$, where CV_k is the cross-validation score, k is the number of folds, and Accuracy_i is the accuracy of the model on the i-th fold. The process began with cross-validation, where each model was trained and evaluated using 5-fold cross-validation. This technique divides the dataset into five equal parts, or folds. In each iteration, one fold is held out as the testing set while the remaining four folds are used for training. This process is repeated five times, with each fold serving as the testing set once. Cross-validation helps mitigate overfitting and provides a more reliable estimate of a model's performance on unseen data by ensuring that each data point is used for both training and validation.

For models that required hyperparameter tuning, GridSearchCV was employed to identify the optimal hyperparameters. GridSearchCV systematically searches through a specified parameter grid, evaluating each combination through cross-validation. The parameter grids were carefully defined based on insights from previous studies and empirical results to ensure a thorough exploration of potential model configurations. This step was crucial in enhancing model performance by fine-tuning parameters such as the number of trees in ensemble methods, the depth of the trees, and learning rates. The objective function for GridSearchCV can be expressed as $\hat{\theta} = \arg\max_{\theta\in\Theta} \frac{1}{k}\sum_{i=1}^{k} \text{Score}(M_\theta, D_i)$, where $\hat{\theta}$ represents the optimal hyperparameters, Θ is the set of all possible hyperparameter values, Score is the evaluation metric (e.g., accuracy, balanced accuracy), M_θ is the model with hyperparameters θ, and D_i is the i-th fold of the dataset.

Once the optimal hyperparameters were identified, each model was trained on the entire training set using these parameters. This ensured that the models were built using the best possible configuration, tailored specifically to the dataset at hand. The training phase involved fitting the models to the data, allowing them to learn the underlying patterns and relationships necessary for making accurate predictions. The evaluation of each model's performance was carried out using the balanced accuracy score. The balanced accuracy score was chosen to address any class imbalance in the dataset by averaging the recall obtained on each class. This metric provides a more balanced view of model performance, ensuring that the model is not biased towards the majority class. The balanced accuracy score was chosen to address any class imbalance in the dataset by averaging the recall obtained on each class. Mathematically, balanced accuracy can be defined as $\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$, where TP is the number of true positives, FN is the number of false negatives, TN is the number of true negatives, and FP is the number of false positives. This metric provides a more balanced view of model performance, ensuring that the model is not biased towards the majority class.

2.7. Performance Metrics

The evaluation of the models in this study was carried out using several key performance metrics designed to provide a comprehensive understanding of each model's effectiveness in diagnosing anemia. These metrics were chosen to account for the nuances of medical datasets, where the costs of false negatives and positives can be significant. The balanced accuracy score was the primary metric used for evaluation. Unlike overall accuracy, which can be misleading in the presence of class imbalance, the balanced accuracy score considers both sensitivity (true positive rate) and specificity (true negative rate). By averaging these two measures, it provides a more equitable assessment of model performance, ensuring that the model's ability to correctly identify both positive and negative cases is accurately reflected. This is particularly important in medical diagnostics, where both types of errors can have serious implications. Cross-validation scores were reported to assess the stability and generalizability of each model. The mean and standard deviation of the cross-validation scores were calculated across the folds used in the K-Fold cross-validation. The mean score provides an estimate of the model's overall performance, while the standard deviation indicates the variability of the performance across different subsets of the data. A low standard deviation suggests that the model performs consistently across different samples, which is indicative of a robust model.

2.8. Hyperparameter Tuning

Hyperparameter tuning was a critical step in optimizing the performance of the selected models. This process was conducted using GridSearchCV, which systematically searches through a predefined set of hyperparameter values to identify the best configuration for each model. The predefined parameter grids were based on both empirical results and insights from previous research, ensuring a thorough exploration of the potential model configurations. For the RandomForestClassifier, the parameter grid included the number of estimators (n_estimators) and the maximum depth of the trees (max_depth). The n_estimators parameter was varied between 100 and 200, while max_depth was tested at values of None (indicating no limit), 10, and 20. This allowed the model to be fine-tuned for both the complexity and the number of decision trees, balancing the trade-off between bias and variance.

The XGBClassifier's parameter grid also included n_estimators and max_depth, with n_estimators set to 100 and 200, and max_depth tested at values of 3, 6, and 10. These parameters control the number of boosting rounds and the depth of each tree, respectively. By optimizing these parameters, the model's ability to fit complex patterns in the data without overfitting was enhanced. For the LGBMClassifier, the parameter grid was similar, with n_estimators set to 100 and 200, and max_depth tested at None, 10, and 20. LightGBM's gradient-based one-sided sampling and exclusive feature bundling techniques were leveraged to ensure efficient training even with these varying parameters. The CatBoostClassifier's hyperparameters included the number of iterations and the depth of the trees. The iterations parameter was set to 100 and 200, while depth was tested at values of 6 and 10. CatBoost's ability to handle categorical features natively without extensive preprocessing was particularly beneficial, and optimizing these parameters helped in enhancing the model's performance and interpretability.

3. RESULTS AND ANALYSIS

This section presents and discusses the results obtained from the training and evaluation of various machine learning models on the anemia diagnosis dataset. The primary objective was to identify the most effective model for accurately classifying the type of anemia based on complete blood count (CBC) data. The performance of each model was assessed using balanced accuracy scores derived from 5-fold cross-validation. The results indicate varying degrees of success across the different models as presented in table. The DecisionTreeClassifier demonstrated remarkable performance with balanced accuracy scores averaging around 94.17%. The scores for this model were consistent across the folds, indicating its reliability and robustness in classifying anemia types. The ExtraTreeClassifier, on the other hand, showed significantly lower performance, with an average balanced accuracy score of approximately 66.3%. This model exhibited considerable variability across the folds, reflecting its sensitivity to the dataset's inherent variability and potential overfitting issues.

The RandomForestClassifier, tuned with a maximum depth of 10 and 100 estimators, achieved a modest average balanced accuracy score of 66.3%. While random forests are generally robust, the selected hyperparameters may have limited the model's complexity, impacting its performance. The ExtraTreesClassifier performed slightly better than its simpler counterpart, the ExtraTreeClassifier, with an average balanced accuracy score of approximately 74.2%. This improvement can be attributed to the ensemble approach, which reduces overfitting by averaging multiple decision trees. The XGBoost model, with a maximum depth of 3 and 100 estimators, achieved an average balanced accuracy score of 74.2%. XGBoost's performance was consistent, and its advanced boosting techniques likely contributed to its relatively high accuracy. Similarly, the LightGBM model, with no maximum depth and 100 estimators, also achieved an average balanced accuracy score of 74.2%. LightGBM's efficient handling of large datasets and complex interactions between features likely contributed to its performance.

The CatBoost model, with a depth of 6 and 200 iterations, performed comparably to LightGBM and XGBoost, with an average balanced accuracy score of 74.2%. CatBoost's unique handling of categorical data and robust boosting algorithms ensured its competitive performance.Among all the models, the DecisionTreeClassifier emerged as the best performer with the highest average balanced accuracy score of

94.17%. The results of this study underscore the effectiveness of simple decision tree models in medical diagnosis tasks, particularly when the dataset is well-preprocessed and balanced. The DecisionTreeClassifier's high performance can be attributed to its ability to capture complex decision boundaries and interactions between features without overfitting.Ensemble methods like RandomForestClassifier, ExtraTreesClassifier, XGBoost, LightGBM, and CatBoost, while generally robust, did not outperform the simple DecisionTreeClassifier in this specific context. This finding suggests that for this particular dataset, the additional complexity introduced by ensemble methods may not provide a significant advantage. It also highlights the importance of hyperparameter tuning and the potential impact of model complexity on performance.

Table 1. The Performance of Machine Learning

Model	Best Params	CV Scores	Mean CV Score
DecisionTree	{}	[0.9896, 0.8869, 0.9819, 0.9574, 0.9924]	0.9616
ExtraTree	{}	[0.5816, 0.7382, 0.7171, 0.6629, 0.6135]	0.6628
RandomForest	{'max_depth': 10, 'n_estimators': 100}	[0.5816, 0.7382, 0.7171, 0.6629, 0.6135]	0.6628
ExtraTrees	{}	[0.6075, 0.7635, 0.8662, 0.7389, 0.7374]	0.7423
XGBoost	{'max_depth': 3, 'n_estimators': 100}	[0.6075, 0.7635, 0.8662, 0.7389, 0.7374]	0.7423
LightGBM	{'max_depth': None, 'n_estimators': 100}	[0.6075, 0.7635, 0.8662, 0.7389, 0.7374]	0.7423
CatBoost	{'depth': 6, 'iterations': 200}	[0.6075, 0.7635, 0.8662, 0.7389, 0.7374]	0.7423

The variability in the performance of the ExtraTreeClassifier indicates that simpler models may struggle with the dataset's complexity, leading to less reliable predictions. This reinforces the need for careful model selection and tuning in medical diagnosis applications. The balanced accuracy scores across different models also highlight the challenge of dealing with imbalanced datasets in medical research. While the DecisionTreeClassifier managed to perform well, other models showed varying degrees of sensitivity to class imbalance, impacting their overall accuracy. Hyperparameter tuning played a crucial role in optimizing the performance of the models. GridSearchCV was employed to identify the best hyperparameters for each model. For instance, the RandomForestClassifier performed best with a maximum depth of 10 and 100 estimators. Similarly, the XGBoost and LightGBM models achieved optimal performance with 100 estimators and specific maximum depth settings. The tuning process underscored the importance of selecting appropriate hyperparameters to balance model complexity and generalization. The selected hyperparameters ensured that the models were neither too simplistic nor overly complex, allowing them to generalize well to unseen data.

4. CONCLUSION

This study aimed to develop and evaluate machine learning models for diagnosing anemia using complete blood count (CBC) data. The results demonstrate the significant potential of machine learning techniques in enhancing the accuracy and efficiency of medical diagnostics, particularly in hematology. The analysis revealed that the DecisionTreeClassifier was the most effective model, achieving the highest balanced accuracy score of 94.17%. This model demonstrated strong performance in classifying anemia types, with minimal false positives and false negatives, indicating its robustness and reliability. The simplicity and interpretability of decision trees make them particularly suitable for clinical applications, where understanding the decision-making process is crucial.

Ensemble methods such as RandomForestClassifier, ExtraTreesClassifier, XGBoost, LightGBM, and CatBoost were also evaluated. While these models are generally powerful and robust, they did not outperform the simpler DecisionTreeClassifier in this specific context. This finding suggests that for this dataset, the additional complexity introduced by ensemble methods may not provide significant advantages. It also underscores the importance of model selection and the impact of hyperparameter tuning on performance. The study also highlighted the importance of addressing class imbalance and optimizing hyperparameters to enhance model performance. Techniques such as GridSearchCV were employed to identify the best hyperparameters, ensuring that the models were neither too simplistic nor overly complex.

A key novelty of this research lies in the comparative analysis of various machine learning models, demonstrating that simpler models like DecisionTreeClassifier can outperform more complex ensemble methods for specific datasets. This insight is valuable for clinical applications where model interpretability is crucial. Overall, this research demonstrates the efficacy of machine learning models in diagnosing anemia based on CBC data. The findings emphasize the potential of these techniques to complement traditional diagnostic methods, providing rapid, automated, and accurate analysis. Future research should explore the integration of more advanced feature engineering, deep learning approaches, and larger, more diverse datasets to further improve model performance and generalizability. Additionally, investigating the application of these models in real-world clinical settings and their impact on diagnostic workflows would provide further validation of their utility.

ACKNOWLEDGEMENTS

We express our deepest gratitude to Atma Jaya Catholic University of Indonesia for the unwavering support and resources provided throughout the duration of this research project. Our sincere thanks also extend to the Information System Study Program, whose rigorous academic environment and collaborative ethos have been instrumental in facilitating our work.

REFERENCES

[1] M. A. Warner and A. C. Weyand, “The Global Burden of Anemia,” in *Blood Substitutes and Oxygen Biotherapeutics*, Springer, 2022, pp. 53–59.

[2] V. Mattiello, M. Schmugge, H. Hengartner, N. von der Weid, R. Renella, and S. P. H. W. Group, “Diagnosis and management of iron deficiency in children with or without anemia: consensus recommendations of the SPOG Pediatric Hematology Working Group,” *Eur. J. Pediatr.*, vol. 179, pp. 527–545, 2020.

[3] D. Kinyoki, A. E. Osgood-Zimmerman, N. V Bhattacharjee, N. J. Kassebaum, and S. I. Hay, “Anemia prevalence in women of reproductive age in low-and middle-income countries between 2000 and 2018,” *Nat. Med.*, vol. 27, no. 10, pp. 1761–1782, 2021.

[4] R. P. B. Tonino, L. M. Zwaginga, M. R. Schipperus, and J. J. Zwaginga, “Hemoglobin modulation affects physiology and patient reported outcomes in anemic and non-anemic subjects: An umbrella review,” *Front. Physiol.*, vol. 14, p. 1086839, 2023.

[5] J. D. Cooper and J. M. Tersak, “Red Blood Cells,” *Zitelli Davis' Atlas Pediatr. Phys. Diagnosis, E-b. Zitelli Davis' Atlas Pediatr. Phys. Diagnosis, E-b.*, vol. 16, no. 13.5, p. 424, 2021.

[6] O. V Chinelo, E. Chukwuka, A. C. Ifeoma, and others, “Causes of anemia due to diminished red blood cell production in pediatrics,” *Int. J. Sci. Adv.*, vol. 3, no. 5, pp. 711–718, 2022.

[7] J. Cotter, C. Baldaia, M. Ferreira, G. Macedo, and I. Pedroto, “Diagnosis and treatment of iron-deficiency anemia in gastrointestinal bleeding: A systematic review,” *World J. Gastroenterol.*, vol. 26, no. 45, p. 7242, 2020.

[8] M. D. Cappellini, K. M. Musallam, and A. T. Taher, “Iron deficiency anaemia revisited,” *J. Intern. Med.*, vol. 287, no. 2, pp. 153–170, 2020.

[9] H. Tvedten, “Classification and laboratory evaluation of anemia,” *Schalm's Vet. Hematol.*, pp. 198–208, 2022.

[10] S. Gajbhiye and J. Aate, “Blood Report Analysis-A Review,” *Trop. J. Pharm. Life Sci.*, vol. 10, no. 5, pp. 63–79, 2023.

[11] K. Delikoyun, E. Cine, M. Anil-Inevi, O. Sarigil, E. Ozcivici, and H. C. Tekin, “2 Deep learning-based cellular image analysis for intelligent medical diagnosis,” in *Artificial Intelligence for Data-Driven Medical Diagnosis*, De Gruyter, 2021, pp. 19–54.

[12] S. A. Wulandari, H. Al Azies, M. Naufal, W. A. Prasetyanto, F. A. Zahra, and others, “Breaking Boundaries in Diagnosis: Non-Invasive Anemia Detection Empowered by AI,” *IEEE Access*, 2024.

[13] M. Sarker, “Revolutionizing healthcare: the role of machine learning in the health sector,” *J. Artif. Intell. Gen. Sci. ISSN 3006-4023*, vol. 2, no. 1, pp. 36–61, 2024.

[14] J. Peng, E. C. Jury, P. Dönnes, and C. Ciurtin, “Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges,” *Front. Pharmacol.*, vol. 12, p. 720694, 2021.

[15] S. Hosseinzadeh Kassani and others, “Towards secure and intelligent diagnosis: deep learning and blockchain technology for computer-aided diagnosis systems,” University of Saskatchewan, 2021.

[16] A. Baldi and S.-R. Pasricha, “Anaemia: Worldwide Prevalence and Progress in Reduction,” in *Nutritional Anemia*, Springer, 2022, pp. 3–17.

[17] M. A. Salam, M. Y. Al-Amin, J. S. Pawar, N. Akhter, and I. B. Lucy, “Conventional methods and future trends in antimicrobial susceptibility testing,” *Saudi J. Biol. Sci.*, vol. 30, no. 3, p. 103582, 2023.

[18] B. R. McFadden, T. J. J. Inglis, and M. Reynolds, “Machine learning pipeline for blood culture outcome prediction using Sysmex XN-2000 blood sample results in Western Australia,” *BMC Infect. Dis.*, vol. 23, no. 1, p. 552, 2023.

[19] J. Mateo, J. M. Rius-Peris, A. I. Maraña-Pérez, A. Valiente-Armero, and A. M. Torres, “Extreme gradient boosting machine learning method for predicting medical treatment in patients with acute bronchiolitis,” *Biocybern. Biomed. Eng.*, vol. 41, no. 2, pp. 792–801, 2021.

[20] S. Pullakhandam and S. McRoy, “Classification and Explanation of Iron Deficiency Anemia from Complete Blood Count Data Using Machine Learning,” *BioMedInformatics*, vol. 4, no. 1, pp. 661–672, 2024.

[21] Y. Wu, L. Zhang, U. A. Bhatti, and M. Huang, “Interpretable machine learning for personalized medical recommendations: A LIME-based approach,” *Diagnostics*, vol. 13, no. 16, p. 2681, 2023.

[22] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, “AI in health and medicine,” *Nat. Med.*, vol. 28, no. 1, pp. 31–38, 2022.

[23] E. Aboelnaga, “Anemia Types Classification.” 2023.

BIBLIOGRAPHY OF AUTHORS

Gregorius Airlangga, Received the B.S. degree in information system from the Yos Sudarso Higher School of Computer Science, Purwokerto, Indonesia, in 2014, and the M.Eng. degree in informatics from Atma Jaya Yogyakarta University, Yogyakarta, Indonesia, in 2016. He got Ph.D. degree with the Department of Electrical Engineering, National Chung Cheng University, Taiwan. He is also an Assistant Professor with the Department of Information System, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia. His research interests include artificial intelligence and software engineering include path planning, machine learning, natural language processing, deep learning, software requirements, software design pattern and software architecture.