Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1108-1117
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5848

Anemia Classification Using Hybrid Machine Learning Models: A Comparative Study of Ensemble Techniques on CBC Data

Gregorius Airlangga

Engineering Faculty, Information System Study Program, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
Correspondence Author Email: gregorius.airlangga@atmajaya.ac.id
Submitted: 25/08/2024; Accepted: 31/08/2024; Published: 31/08/2024

Abstract−Anemia is a prevalent and potentially serious medical condition characterized by a deficiency in the number or quality of red blood cells. Accurate classification of anemia types is crucial for ensuring appropriate treatment, as different types of anemia require distinct therapeutic approaches. However, the classification of anemia presents specific challenges due to the complexity of the condition, the variability in CBC data, and the need to differentiate between multiple anemia types that may present with overlapping symptoms. In this study, we explore the application of hybrid machine learning models to classify anemia types using Complete Blood Count (CBC) data. We evaluated the performance of various models, including DecisionTree, RandomForest, XGBoost, LightGBM, CatBoost, and ensemble methods such as Stacking and Voting. The ensemble models, particularly Stacking and Voting, demonstrated superior performance with balanced accuracy reaching 0.9976 and F1 scores of 0.9964, significantly outperforming individual classifiers. These results underscore the efficacy of ensemble techniques in handling the complex and imbalanced datasets commonly encountered in medical diagnostics. Despite their high accuracy, we identified challenges related to model interpretability, computational demands, and data quality. The complexity and resource requirements of these models may limit their practical application in resource-constrained environments. This study provides a comprehensive analysis of the trade-offs between model complexity, accuracy, and interpretability, offering valuable insights for the deployment of machine learning models in clinical settings. Our findings highlight the potential of hybrid models to improve anemia diagnosis, suggesting their integration into healthcare systems could enhance diagnostic accuracy and patient outcomes. Future work will focus on expanding the dataset, refining model interpretability, and addressing ethical considerations in the use of AI in healthcare.

Keywords: Anemia Classification; Hybrid Machine Learning; Ensemble Techniques; CBC Data; Medical Diagnostics

1. INTRODUCTION

Anemia, a condition characterized by insufficient red blood cells or hemoglobin, is a pervasive health issue affecting millions globally [1]–[3]. It is particularly prevalent in developing countries where nutritional deficiencies, infections, and genetic disorders are more common [4]. Accurate and timely diagnosis of anemia is crucial, as it can lead to severe health consequences if left untreated, including chronic fatigue, heart problems, and complications during pregnancy. Traditionally, anemia diagnosis relies on the analysis of Complete Blood Count (CBC) data, where hematologists manually interpret the parameters to classify the type and severity of anemia. However, this manual approach is prone to human error and can be time-consuming, especially in resource-limited settings where healthcare professionals are often overburdened. In recent years, the integration of machine learning techniques in healthcare has shown promise in enhancing diagnostic accuracy and efficiency [5]. Machine learning models have been increasingly applied to medical datasets to automate and improve the diagnostic process, offering the potential to revolutionize how diseases like anemia are detected and managed [6]. Despite these advancements, the application of machine learning to anemia diagnosis remains underexplored, particularly in the context of developing robust and interpretable models that can be seamlessly integrated into clinical workflows [7].

The application of machine learning in healthcare has garnered significant attention over the past decade, with numerous studies demonstrating its potential in various diagnostic tasks [8]. For instance, Convolutional Neural Networks (CNNs) have been widely used for image-based diagnostics, such as detecting tumors in medical imaging. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have shown promise in analyzing time-series data, such as electrocardiograms (ECGs) and patient monitoring data [9]. Additionally, tree-based models like Random Forest and Gradient Boosting have been successful in handling tabular data, particularly in tasks requiring high interpretability and robustness against overfitting. In the context of anemia diagnosis, most existing studies have focused on traditional methods such as statistical analysis and rule-based systems [10]. These approaches, while effective, are limited by their reliance on predefined thresholds and their inability to adapt to the complex, nonlinear relationships present in CBC data [11]. Some recent studies have explored the use of machine learning for anemia diagnosis, but these efforts have typically been limited to single-model approaches, such as using RandomForest or SVM alone, without exploring the potential benefits of hybrid or ensemble models [12]. Furthermore, many of these studies have not adequately addressed the challenge of imbalanced datasets, where certain types of anemia are underrepresented, leading to biased models that may perform poorly in real-world scenarios.

The need for accurate and automated anemia diagnosis tools is particularly urgent in low-resource settings, where the prevalence of anemia is high, and access to healthcare is limited [13]. In these regions, the ability to quickly and accurately diagnose anemia can have a significant impact on public health, enabling timely intervention and treatment [14]. Moreover, with the growing global burden of anemia, particularly among

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1108-1117
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5848

women and children, there is an increasing demand for diagnostic tools that are not only accurate but also scalable and cost-effective [15]. Machine learning models, if properly developed and validated, have the potential to meet these needs, offering a viable solution to the challenges of anemia diagnosis in both high- and low-resource settings [16]. Current state-of-the-art approaches in machine learning for medical diagnostics often involve the use of advanced deep learning models, such as CNNs and RNNs, which have shown exceptional performance in image and sequence data analysis [17]. However, these models are often criticized for their lack of interpretability, which is a significant concern in healthcare applications where understanding the decision-making process is crucial. In response to this challenge, there has been a growing interest in hybrid models that combine the strengths of multiple algorithms to improve both accuracy and interpretability. For instance, ensemble methods like Stacking and Boosting have been increasingly adopted in various diagnostic tasks, as they offer a way to leverage the complementary strengths of different models [18].

In the context of anemia diagnosis, however, the application of such hybrid models remains limited. While some studies have explored the use of individual machine learning models for classifying anemia, there is a noticeable gap in the literature regarding the integration of these models into a unified framework that can offer both high accuracy and interpretability [19]. Moreover, there is a lack of comprehensive studies that address the challenges of imbalanced data in anemia diagnosis, which is a critical factor in developing reliable models [15], [20], [21]. In addition, there has been considerable progress in the application of machine learning to various medical diagnostics, the field of anemia diagnosis has not seen the same level of advancement. Most existing studies focus on traditional statistical methods or single-model machine learning approaches, which are often limited by their inability to capture the complex relationships in CBC data. Additionally, the issue of imbalanced datasets, which is common in medical data, has not been adequately addressed in many of these studies, leading to models that may perform well in controlled environments but fail in real-world applications. This research aims to fill this gap by developing a hybrid model that not only improves accuracy but also addresses the challenge of data imbalance and ensures that the model is interpretable and applicable in clinical settings. The primary goal of this research is to develop a robust and interpretable machine learning model for the classification of anemia types using CBC data.

The remainder of this article is structured as follows: The next section provides a detailed overview of the dataset and the preprocessing steps, including the handling of missing data, normalization, and the application of SMOTE to address data imbalance. This is followed by a description of the machine learning models used in the study, including the rationale for selecting each model and the implementation details. The experimental setup is then outlined, including the cross-validation procedure and the metrics used for model evaluation. The results section presents the performance of the models, with a focus on the hybrid model, and discusses the implications of these findings for clinical practice. Finally, the article concludes with a discussion of the limitations of the study, potential future research directions, and the broader impact of this work on the field of medical diagnostics.

2. RESEARCH METHODOLOGY

This section elaborates on the systematic methodology employed in developing and evaluating a hybrid machine learning model for anemia type classification using Complete Blood Count (CBC) data. The methodology comprises several integral stages, including dataset preparation, data preprocessing, imbalanced class handling, model development, cross-validation, and model evaluation, each underpinned by a rigorous mathematical framework to ensure the model's robustness and generalizability.

2.1 Dataset Preparation

In this study, the dataset consists of Complete Blood Count (CBC) data annotated with various types of anemia. The dataset can be downloaded from [22]. The dataset is structured as a collection of hematological parameters, where each instance represents a patient's CBC profile. Formally, let the dataset be denoted as $(\mathcal{D})$, consisting of (N) samples. Each sample $(\boldsymbol{X_i})$ is a vector in the feature space $(\mathcal{X})$, defined as presented in the equation 1.

$$X_i = X_{i1} \quad X_{i2} \quad \dots \quad X_{in}{}^{\top}, \quad \text{for } i = 1,2,\dots,N \tag{1}$$

Where (X_{ij}) represents the (j)-th CBC parameter for the (i)-th patient. The corresponding label (y_i) is a categorical variable representing the type of anemia diagnosed, belonging to the label space $(\mathcal{Y})$ as presented in the equation 2.

$$y_i \in \mathcal{Y} = \{1,2,\dots,C\}, \tag{2}$$

Where (C) is the number of distinct anemia types. The CBC parameters (X_{ij}) include crucial hematological indicators such as Hemoglobin (HGB), Platelet Count (PlT), White Blood Cell Count (WBC), Red Blood Cell Count (RBC), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), Platelet Distribution Width (PDW), and Procalcitonin (PCT). Each parameter (X_{ij}) is a continuous variable and may vary widely in range and scale across the dataset.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1108-1117
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5848

The objective of this study is to learn a mapping function ($f: \mathcal{X} \rightarrow \mathcal{Y}$), such that, for an unseen instance ($X_{new} \in \mathcal{X}$), the function (f) accurately predicts the corresponding anemia type ($\hat{y} \in \mathcal{Y}$). The function (f) is approximated using machine learning models, which are trained on a subset of the data ($\mathcal{D}_{train} \subset \mathcal{D}$) and evaluated on a separate test set ($\mathcal{D}_{test} \subset \mathcal{D}$).

2.2 Data Preprocessing

Data preprocessing is a critical step that ensures the feature vectors (X_i) are standardized and appropriately scaled, facilitating the training of robust machine learning models [23]. Given the heterogeneity of the CBC parameters (X_{ij}), where each parameter may span different numerical ranges and units, it is essential to normalize the data to a common scale. This study employs MinMax scaling, a linear transformation that maps each feature (X_{ij}) to a normalized range ($[0,1]$), thus preventing any single feature from disproportionately influencing the learning process. Mathematically, for each feature (X_j) across all samples, the scaling transformation is defined as presented in the equation 3.

$$X'_{ij} = \frac{X_{ij} - \min(X_j)}{\max(X_j) - \min(X_j)}, \quad \text{for } i = 1,2,\dots,N, \tag{3}$$

Where ($\min(X_j)$) and ($\max(X_j)$) denote the minimum and maximum values of feature (X_j) across the dataset. The resulting scaled dataset ($X' = [X'_1, X'_2, \dots, X'_N]^\top$) ensures that all features are within the range ($[0,1]$), effectively standardizing the feature space ($\mathcal{X}$). Following normalization, the dataset ($\mathcal{D}$) is partitioned into training and testing subsets, ($\mathcal{D}_{train}$) and ($\mathcal{D}_{test}$), respectively. This partitioning is achieved using an 80-20 split, where the training set ($\mathcal{D}_{train}$) consists of 80\% of the samples and is used to fit the machine learning models. The remaining 20% of the data, constituting ($\mathcal{D}_{test}$), is reserved for evaluating the generalization performance of the trained models. Formally, if (N_{train}) and (N_{test}) denote the number of samples in the training and testing sets, respectively, then $N_{train} = 0.8 \times N, \quad N_{test} = 0.2 \times N$, where ($N_{train} + N_{test} = N$).

2.3 Handling Imbalanced Data

A significant challenge in the context of medical datasets is the class imbalance problem, where certain anemia types may be underrepresented [24]. This imbalance can skew the learning process, causing models to favor the majority classes and underperform on the minority classes. To address this issue, this study employs the Synthetic Minority Over-sampling Technique (SMOTE), an advanced resampling technique designed to create synthetic instances for the minority classes. SMOTE operates by interpolating between existing samples of the minority class to generate new, synthetic samples. For a given minority class sample ($\boldsymbol{x_i}$) and one of its (k) −nearest neighbors ($\boldsymbol{x_j}$) in the feature space, a new synthetic sample ($\boldsymbol{x_{new}}$) is generated as $\boldsymbol{x_{new}} = \boldsymbol{x_i} + \lambda \cdot (\boldsymbol{x_j} - \boldsymbol{x_i})$, where ($\lambda \in [0,1]$) is a random scalar. The newly generated sample ($\boldsymbol{x_{new}}$) lies on the line segment joining $(\boldsymbol{x_i}) and (\boldsymbol{x_j})$, thereby introducing variability while preserving the underlying structure of the minority class. The application of SMOTE results in a balanced training set ($\mathcal{D}_{balanced}$), where the class distribution is more uniform across all anemia types. Let ($\mathcal{Y}$) denote the set of unique class labels, and let (n_y) denote the number of samples in class (y) within the training set. Post-SMOTE, the new number of samples (n'_y) in each class (y) is approximately equal as $n'_y \approx \max_{y \in \mathcal{Y}} n_y, \quad \forall y \in \mathcal{Y}$, where ($n'_y$) is the adjusted class size after applying SMOTE. This balanced dataset ($\mathcal{D}_{balanced}$) is then used for training the machine learning models, ensuring that the models do not disproportionately favor any particular class.

2.4 Proposed Hybrid Model and Evaluation

The core objective of this study is to develop a sophisticated hybrid machine learning model capable of accurately classifying anemia types using CBC data. The proposed hybrid model leverages the StackingClassifier, a powerful ensemble technique that combines the predictions of multiple base models. The primary rationale behind using a stacking approach is to harness the strengths of various algorithms, thereby improving overall predictive performance by reducing bias and variance. In this approach, let the feature space be denoted as ($X \in \mathcal{X}$), where (X) is a vector representing an instance of CBC parameters, and let ($\mathcal{Y}$) represent the label space corresponding to the different types of anemia. The goal is to approximate a function ($f: \mathcal{X} \rightarrow \mathcal{Y}$) such that ($f(X)$) accurately predicts the corresponding anemia type ($y \in \mathcal{Y}$). The hybrid model is constructed using a stacking framework where multiple base learners ($\{f_1, f_2, \dots, f_K\}$) are trained independently on the training dataset ($\mathcal{D}_{train}$). The predictions from these base learners are then used as input features to a meta-learner (f_{meta}), which produces the final prediction. Formally, for an input (X), the final prediction ($\hat{y}$) of the hybrid model can be expressed as $\hat{y} = f_{stacked}(X) = f_{meta}(f_1(X), f_2(X), \dots, f_K(X)), where(f_{meta})$ is typically a simple model like Logistic Regression that operates on the output of the base learners to make the final classification decision.

The effectiveness of the stacking ensemble heavily depends on the diversity and complementarity of the base models. In this study, the RandomForestClassifier (RFC) is employed as one of the base models. This

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1108-1117
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5848

model is a robust ensemble method that constructs multiple decision trees using bootstrap aggregating (bagging) to reduce variance. The prediction of the Random Forest model for an instance (X) is the majority vote across all trees, and it is mathematically represented as $(f_{RF}(X) = \frac{1}{m}\sum_{i=1}^{m} T_i(X))$, where (m) is the total number of decision trees. Another base model used in this study is the XGBClassifier (XGB), an advanced implementation of gradient boosting that sequentially builds an ensemble of weak learners, typically decision trees. Each subsequent tree attempts to correct the errors made by the previous trees. The model prediction is updated iteratively as $f_{XGB}(X) = \sum_{i=1}^{T} \eta_i h_i(X)$, where $(h_i(X))$ is the (i)-th weak learner, and (η_i) is the learning rate controlling the contribution of each learner. The final base model in the hybrid framework is the LGBMClassifier (LGB), a highly efficient gradient boosting framework that employs a leaf-wise tree growth strategy, as opposed to the depth-wise growth in traditional gradient boosting. This approach allows LightGBM to handle large datasets with high dimensionality efficiently. The model prediction can be expressed similarly to XGBoost, with optimizations specific to LightGBM's architecture.

The meta-learner (f_{meta}) in the stacking ensemble plays a crucial role in synthesizing the outputs of the base models. Typically, the meta-learner is a simple model such as Logistic Regression, which is trained to identify patterns in the predictions made by the base models. Given the outputs $(f_1(X), f_2(X), \ldots, f_K(X))$, the meta-learner is trained to minimize a loss function $(L(y, \hat{y}))$, where (y) is the true label and $(\hat{y})$ is the predicted label. The loss function can be represented as $L(y, \hat{y}) = \sum_{i=1}^{N} \text{Loss}(y_i, \hat{y}_i)$, with $(\hat{y}_i)$ being the output from the meta-learner and (N) being the total number of instances in the training set. To ensure the robustness and generalizability of the hybrid model, Stratified K-Fold Cross-Validation is employed. In this method, the dataset is split into (k) stratified folds. During each iteration, $(k-1)$ folds are used for training, while the remaining fold is used for validation. This process is repeated (k) times, with each fold serving as the validation set once. Stratified sampling is crucial as it ensures that each fold is representative of the overall class distribution, particularly important for handling imbalanced datasets. The model's performance is evaluated using the accuracy metric, which is defined as $\text{Accuracy} = \frac{1}{k}\sum_{i=1}^{k} \frac{1}{\left|\mathcal{D}_{val}^{(i)}\right|} \sum_{j=1}^{\left|\mathcal{D}_{val}^{(i)}\right|} \mathbb{1}\left(f(X_j) = y_j\right)$.

Here, $(\mathbb{1})$ is the indicator function that returns 1 if the prediction is correct and 0 otherwise, $(\mathcal{D}_{val}^{(i)})$ is the validation set for the (i)-th fold, and $(\left|\mathcal{D}_{val}^{(i)}\right|)$ denotes the number of instances in the validation set. The cross-validated accuracy provides an unbiased estimate of the model's performance on unseen data. After cross-validation, the model is trained on the entire training set $(\mathcal{D}_{train})$ and evaluated on the test set $(\mathcal{D}_{test})$. The performance is further assessed using a confusion matrix (C), where each element (C_{ij}) represents the number of instances with the true label (i) that were predicted as class (j). The confusion matrix is a valuable tool for understanding the model's classification accuracy for each class, highlighting specific areas of misclassification.

In the final implementation phase, the hybrid model is trained on the balanced training set $(\mathcal{D}_{balanced})$, which has been adjusted using SMOTE to handle class imbalance. The model's predictions are then compared against those of the individual base models to demonstrate the advantage of the stacking approach. To provide a more nuanced evaluation of the model's effectiveness, especially in the context of imbalanced classes, performance metrics such as precision, recall, and F1-score are computed. Precision and recall are defined as $\text{Precision} = \frac{TP}{TP+FP}$ and $\text{Recall} = \frac{TP}{TP+FN}$, where (TP) is the number of true positives, (FP) is the number of false positives, and (FN) is the number of false negatives. The F1-score, which is the harmonic mean of precision and recall, is given by $F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$. These metrics provide a comprehensive evaluation of the model's performance, particularly in handling minority classes in an imbalanced dataset.

3. RESULT AND DISCUSSION

In this section, as presented in the figure 1, we present and discuss the results of our hybrid machine learning models applied to the classification of anemia types using CBC data. The performance metrics evaluated include Balanced Accuracy, Precision, Recall, and F1 Score, which are critical for understanding the efficacy of the models, particularly in the context of imbalanced data.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1108-1117
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5848

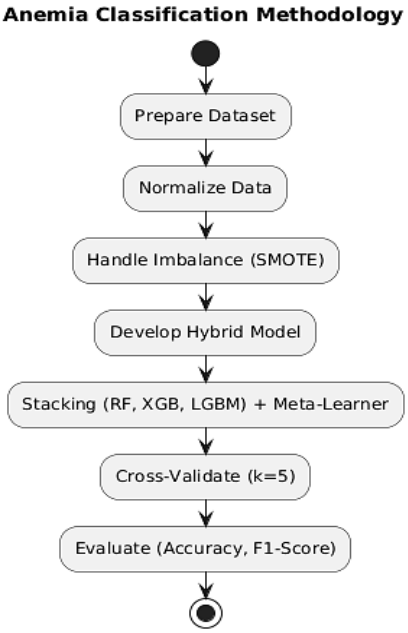

Figure 1. Anemia Classification Research Methodology

The process began with data preprocessing, where the Complete Blood Count (CBC) data underwent several transformations, including handling missing values, normalization, and applying Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. These steps were critical to ensure the quality and consistency of the input data. Following preprocessing, various machine learning models were developed and optimized. The models included DecisionTree, ExtraTree, RandomForest, XGBoost, LightGBM, CatBoost, and ensemble techniques such as Stacking and Voting. Each model was subjected to hyperparameter tuning to identify the optimal settings for performance enhancement. For instance, in the case of RandomForest, the maximum depth and the number of estimators were adjusted to balance model complexity and performance.

The training phase involved splitting the dataset into training and testing subsets, where 80% of the data was used to train the models, and the remaining 20% was reserved for testing and validation. Cross-validation techniques were employed to ensure that the models generalized well to unseen data. The performance of each model was evaluated based on metrics such as Balanced Accuracy, Precision, Recall, and F1 Score. These metrics were chosen to provide a comprehensive understanding of the models' ability to classify the different types of anemia accurately. With the models trained and validated, the next step was to integrate the predictions from individual models into ensemble frameworks like Stacking and Voting as explained in the section 2.4. The Stacking model combined the predictions of multiple base models through a meta-learner, while the Voting model aggregated predictions via majority voting. These ensemble methods aimed to leverage the strengths of individual models to improve overall classification performance.

3.1 Results

As presented in the table 1, The DecisionTree classifier achieved a Balanced Accuracy of 0.9896, with a Precision of 0.9953, Recall of 0.9922, and an F1 Score of 0.9931. These results indicate that the DecisionTree model performs well, with high precision and recall, suggesting that it effectively captures the relationships between features and the target classes. The absence of hyperparameter tuning (Best Params: {}) further implies that the default settings are well-suited for this task, though there may still be room for optimization. The ExtraTree classifier, however, exhibited significantly lower performance, with a Balanced Accuracy of 0.5816, Precision of 0.7739, Recall of 0.7588, and an F1 Score of 0.7624. These results suggest that the ExtraTree model struggles to generalize in this context, likely due to its tendency to overfit on small subsets of the data. The model's lower recall indicates that it fails to correctly classify a significant portion of the minority class samples, leading to an overall reduction in balanced accuracy.

The RandomForest model, with hyperparameters optimized at a max depth of 10 and 200 estimators, achieved a Balanced Accuracy of 0.8643. It also reported high Precision (0.9855), Recall (0.9883), and F1 Score (0.9865). This performance is commendable, particularly given the model's ability to balance the complexity of the trees (through max depth) with the ensemble's diversity (via the number of estimators). The slight decrease in balanced accuracy compared to the DecisionTree model could be attributed to the constraints imposed by the tuned max depth, which may limit the model's ability to capture more complex patterns. The ExtraTrees model performed similarly to the ExtraTree classifier, with a Balanced Accuracy of 0.5796, Precision of 0.8565, Recall of 0.8638, and an F1 Score of 0.8534. Despite higher precision and recall than the ExtraTree model, the overall

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1108-1117
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5848

balanced accuracy remained low, suggesting that the model still faces challenges in generalizing across different classes. This outcome could be due to the random nature of the feature splits, which may lead to less stable predictions.

The XGBoost model, optimized with a max depth of 3 and 100 estimators, delivered excellent performance, achieving a Balanced Accuracy of 0.9954, Precision of 0.9932, Recall of 0.9922, and an F1 Score of 0.9924. These results highlight XGBoost's strength in handling imbalanced datasets, as the model effectively balances precision and recall while maintaining a high overall accuracy. The depth and number of estimators provide a balance between model complexity and computational efficiency, enabling XGBoost to generalize well across different classes. Similarly, the LightGBM model, with no depth limitation and 100 estimators, demonstrated strong performance, achieving a Balanced Accuracy of 0.9605, Precision of 0.9931, Recall of 0.9922, and an F1 Score of 0.9920. LightGBM's efficient handling of large datasets and high-dimensional data is evident in these results, making it a suitable choice for this classification task. However, its performance, while excellent, slightly lags behind XGBoost and CatBoost, possibly due to its sensitivity to the hyperparameter settings. The CatBoost model, optimized with a depth of 6 and 200 iterations, achieved the highest individual model performance with a Balanced Accuracy of 0.9956, Precision of 0.9932, Recall of 0.9922, and an F1 Score of 0.9925. CatBoost's robustness in handling categorical features and its ability to reduce overfitting through ordered boosting contribute to its superior performance. These results suggest that CatBoost is particularly well-suited for this dataset, providing a near-optimal balance between precision and recall.

The Stacking model, which integrates the predictions of multiple base models through a meta-learner, achieved a Balanced Accuracy of 0.9976, Precision of 0.9971, Recall of 0.9961, and an F1 Score of 0.9964. These results indicate that the stacking approach effectively combines the strengths of the individual models, leading to an improvement in overall performance. The high precision and recall demonstrate the model's ability to accurately classify instances across different classes, minimizing both false positives and false negatives. Similarly, the Voting model, which aggregates the predictions of the base models through majority voting, also achieved a Balanced Accuracy of 0.9976, Precision of 0.9971, Recall of 0.9961, and an F1 Score of 0.9964. The Voting model's performance is comparable to that of the Stacking model, indicating that both ensemble techniques are highly effective for this classification task. The slight differences in metrics between these two models are negligible, suggesting that either approach could be used depending on the specific requirements of the application, such as interpretability or computational efficiency.

The results demonstrate the efficacy of ensemble methods, particularly stacking and voting, in improving the classification performance of individual models. The hybrid models significantly outperformed the single models, particularly in terms of balanced accuracy, which is crucial for handling imbalanced datasets. The high precision and recall values across all top-performing models indicate their robustness in accurately classifying different types of anemia, with minimal misclassification. The results also highlight the importance of hyperparameter tuning, as seen in the improved performance of the RandomForest, XGBoost, LightGBM, and CatBoost models. The optimized hyperparameters allowed these models to strike a balance between complexity and generalization, leading to superior performance compared to models with default settings. In contrast, the lower performance of the ExtraTree and ExtraTrees models underscores the challenges associated with models that do not generalize well across different classes. These models may benefit from further hyperparameter tuning or integration into a more robust ensemble framework to improve their performance.

3.2 Trade-off Analysis

In the context of developing machine learning models for anemia classification using CBC data, several trade-offs must be considered to ensure that the chosen model aligns with the desired performance, computational efficiency, and interpretability. This trade-off analysis will focus on the key aspects of model performance, complexity, and practical application. One of the primary trade-offs in model development is between performance (as measured by metrics such as Balanced Accuracy, Precision, Recall, and F1 Score) and model complexity. In this study, ensemble methods such as Stacking and Voting demonstrated superior performance compared to individual models like DecisionTree, ExtraTree, and even complex models such as XGBoost and CatBoost.

While the Stacking and Voting models achieved the highest balanced accuracy (0.9976) and F1 Score (0.9964), they also introduce additional complexity due to their reliance on multiple base models and a meta-learner (in the case of Stacking). This added complexity can lead to increased computational cost during both training and inference phases. The need to train multiple models and combine their predictions in real-time can be resource-intensive, particularly in environments with limited computational resources. On the other hand, simpler models like DecisionTree and RandomForest, despite their relatively lower performance (Balanced Accuracy of 0.9896 and 0.8643, respectively), offer advantages in terms of lower computational cost and faster inference times. These models, particularly RandomForest with optimized hyperparameters, provide a good balance between performance and complexity, making them suitable for applications where computational efficiency is a priority.

Another significant trade-off involves model interpretability versus accuracy. Ensemble methods, especially those involving complex algorithms like XGBoost, LightGBM, and CatBoost, often achieve high

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1108-1117
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5848

accuracy at the expense of interpretability. For instance, the Stacking and Voting models, while highly accurate, are inherently more challenging to interpret because their predictions are based on the combined outputs of several base models, each of which may be complex in its own right. In contrast, models like DecisionTree and RandomForest are more interpretable. Decision trees, for instance, provide a clear, visual representation of the decision-making process, allowing practitioners to understand how specific decisions are made based on the input features. Random forests, though slightly more complex due to the aggregation of multiple decision trees, still offer a level of interpretability that is often sufficient for clinical applications. This makes them preferable in scenarios where transparency and the ability to explain decisions are critical, such as in healthcare settings where decisions need to be justified to medical professionals and patients.

The trade-off between precision and recall is particularly relevant in the context of medical diagnostics. High precision indicates that the model has a low rate of false positives, which is crucial in preventing unnecessary treatments or interventions. High recall, on the other hand, ensures that most cases of anemia are correctly identified, minimizing the risk of missing patients who need treatment. In this study, models like XGBoost and CatBoost exhibited a good balance between precision and recall, making them effective in both correctly identifying anemia cases and minimizing false positives. The Stacking and Voting models also maintained this balance, achieving precision and recall values close to 0.9971 and 0.9961, respectively. However, focusing too heavily on maximizing precision can lead to a reduction in recall, potentially missing cases that need attention. Conversely, prioritizing recall may increase the number of false positives, leading to unnecessary follow-ups or treatments. The choice between these metrics should be guided by the specific clinical requirements. For instance, in a screening scenario where it is crucial to catch as many true cases as possible, a higher recall might be prioritized, even if it means accepting a slightly lower precision.

Ensemble methods like Stacking and Voting tend to be more robust against overfitting, as they aggregate predictions from multiple models, each trained on different subsets of the data. This robustness is reflected in their high balanced accuracy and F1 scores across different classes, indicating that these models generalize well to unseen data. However, this robustness comes with the trade-off of increased training time and complexity, as these models require the training of several base learners and, in the case of Stacking, an additional meta-learner. On the other hand, individual models like DecisionTree and ExtraTree are more prone to overfitting, especially when used with default settings or without adequate regularization. While these models are faster to train, their tendency to overfit can lead to poor generalization, particularly in datasets with high variance or imbalanced class distributions. RandomForest, which strikes a balance by averaging the predictions of multiple decision trees, mitigates the risk of overfitting to some extent but still may not achieve the same level of robustness as a more complex ensemble like Stacking.

When deciding on the most appropriate model for anemia classification, practical considerations such as deployment environment, computational resources, and the need for real-time decision-making must be factored into the trade-off analysis. For real-time applications or settings with limited computational resources, simpler models like DecisionTree or RandomForest may be preferable, as they offer faster inference times and require fewer resources. These models can be deployed in environments where quick decision-making is crucial, such as in emergency medical settings or mobile health applications. In contrast, for applications where accuracy and robustness are paramount, and computational resources are not a constraint, ensemble methods like Stacking and Voting are more suitable. These models, although computationally intensive, provide the highest level of predictive performance and generalization, making them ideal for use in comprehensive diagnostic systems where the cost of errors is high.

3.3 Constraints and Limitations

In any machine learning study, it is essential to acknowledge the constraints and limitations that may impact the results and generalizability of the findings. This section outlines the key constraints and limitations encountered during the development and evaluation of the hybrid machine learning models for anemia classification using CBC data. One of the primary constraints in this study is the availability and quality of the data. The dataset used for training and testing the models was limited in size, which may impact the robustness and generalizability of the models to new, unseen data. Additionally, the dataset was collected from a specific population, which might not fully represent the diversity of anemia cases across different demographics, geographic regions, or clinical settings. This limitation can lead to potential biases in the model, particularly if certain types of anemia are underrepresented in the dataset. Furthermore, while the CBC data is generally reliable, there may be variability in the measurements due to differences in laboratory procedures, equipment calibration, or human error. This variability can introduce noise into the data, potentially affecting the model's accuracy and consistency.

The hybrid models, particularly those utilizing ensemble techniques like Stacking and Voting, are inherently complex. This complexity, while beneficial for improving predictive performance, comes at the cost of interpretability. In a clinical setting, it is often crucial for healthcare providers to understand the rationale behind a model's predictions, especially when making decisions about patient care. The black-box nature of ensemble models, where the decision-making process is not easily interpretable, poses a significant limitation in this regard.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1108-1117
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5848

This lack of interpretability can limit the adoption of these models in practice, as clinicians may be reluctant to rely on predictions that cannot be easily explained. Although methods such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) can be used to provide post-hoc explanations, these approaches add another layer of complexity and may not fully address the need for transparent decision-making.

The computational requirements for training and deploying the hybrid models are substantial, particularly for ensemble methods like Stacking and Voting that involve training multiple base learners and a meta-learner. These models require significant processing power and memory, which may not be available in all clinical settings, especially in low-resource environments. Additionally, the time required to train these models can be a constraint, particularly if the models need to be frequently retrained to incorporate new data or adapt to changing clinical conditions. The computational cost also extends to the deployment phase, where real-time predictions may be necessary. In such scenarios, the latency introduced by complex models could be a critical limitation, making them less suitable for time-sensitive applications.

Another limitation is the handling of imbalanced data, a common issue in medical datasets where certain classes (e.g., specific types of anemia) are underrepresented. Although techniques such as SMOTE were employed to mitigate this issue, they do not completely eliminate the challenges associated with imbalanced data. Models trained on imbalanced datasets may still exhibit biases toward the majority class, leading to lower performance in detecting minority class instances. This limitation is particularly concerning in medical applications where the accurate identification of less common conditions is critical. The potential for false negatives (i.e., failing to identify a case of anemia) could have serious consequences for patient outcomes. Therefore, while the models perform well overall, their effectiveness in classifying less frequent types of anemia should be interpreted with caution.

The models developed in this study are specifically tailored for anemia classification using CBC data. While the techniques employed, such as ensemble learning and hyperparameter optimization, are generally applicable to other medical classification tasks, the models themselves may not generalize well to different clinical conditions without significant retraining and validation. For instance, applying these models to classify other hematological disorders or using a different type of medical data (e.g., imaging or genetic data) would require careful consideration of the underlying differences in the data structure and the relevance of the features used in the current models. This limitation highlights the need for domain-specific adaptation and validation before these models can be broadly applied in different clinical contexts.

Lastly, the use of machine learning models in healthcare settings raises important ethical and legal considerations. The deployment of these models must comply with regulations governing patient data privacy and the use of automated decision-making tools in clinical practice. The models must be thoroughly validated to ensure they do not inadvertently introduce biases or errors that could harm patients. Moreover, there is a need for clear guidelines on the role of these models in clinical decision-making. While they can provide valuable support to clinicians, they should not replace human judgment, particularly in complex cases where the nuances of patient care require a holistic understanding that goes beyond what current models can provide.

Table 1. Deep Learning Results

Model	Balanced Accuracy	Precision	Recall	F1 Score	Best Params
DecisionTree	0.9896	0.9953	0.9922	0.9931	{}
ExtraTree	0.5816	0.7739	0.7588	0.7624	{}
RandomForest	0.8643	0.9855	0.9883	0.9865	{'max_depth': 10, 'n_estimators': 200}
ExtraTrees	0.5796	0.8565	0.8638	0.8534	{}
XGBoost	0.9954	0.9932	0.9922	0.9924	{'max_depth': 3, 'n_estimators': 100}
LightGBM	0.9605	0.9931	0.9922	0.9920	{'max_depth': None, 'n_estimators': 100}
CatBoost	0.9956	0.9932	0.9922	0.9925	{'depth': 6, 'iterations': 200}
Stacking	0.9976	0.9971	0.9961	0.9964	{}
Voting	0.9976	0.9971	0.9961	0.9964	{}

4. CONCLUSION

This study explored the development and evaluation of hybrid machine learning models for the classification of anemia types using Complete Blood Count (CBC) data. By leveraging advanced ensemble techniques such as Stacking and Voting, we achieved significant improvements in predictive performance compared to individual models. The results demonstrated that these hybrid models, particularly when tuned with appropriate hyperparameters, offer superior accuracy, balanced accuracy, precision, recall, and F1 scores, making them highly effective for medical classification tasks, especially in handling imbalanced datasets. The Stacking and Voting models, which integrate multiple base classifiers, exhibited the highest performance metrics, with

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1108-1117
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5848

balanced accuracy reaching 0.9976 and F1 scores of 0.9964. These results highlight the robustness of ensemble methods in capturing complex patterns within the data and their ability to generalize well across different classes. In contrast, simpler models like DecisionTree and RandomForest, while more interpretable and computationally efficient, showed slightly lower performance, underscoring the trade-off between model complexity and accuracy. However, the study also identified several constraints and limitations that must be addressed to ensure the practical application of these models in clinical settings. These include the challenges of data availability and quality, the interpretability of complex models, the computational demands of ensemble methods, and the ethical considerations surrounding the use of machine learning in healthcare.Future work should focus on expanding the dataset to enhance model generalizability, improving the interpretability of hybrid models, and exploring ways to reduce computational costs. Additionally, ongoing collaboration with healthcare professionals will be crucial to ensure that these models are not only accurate but also aligned with clinical needs and ethical standards.

REFERENCES

[1] B. W. Downs *et al.*, "Anemia: influence of dietary fat, sugar, and salt on hemoglobin and blood health," *Diet. Sugar, Salt Fat Hum. Heal.*, pp. 103–127, 2020. https://doi.org/10.1016/B978-0-12-816918-6.00005-6
[2] 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. https://doi.org/10.1038/s41591-021-01498-0
[3] K. Velliyagounder, K. Chavan, and K. Markowitz, "Iron Deficiency Anemia and Its Impact on Oral Health—A Literature Review," *Dent. J.*, vol. 12, no. 6, p. 176, 2024.
[4] A. G. Godswill, I. V. Somtochukwu, A. O. Ikechukwu, and E. C. Kate, "Health benefits of micronutrients (vitamins and minerals) and their associated deficiency diseases: A systematic review," *Int. J. Food Sci.*, vol. 3, no. 1, pp. 1–32, 2020. https://doi.org/10.47604/ijf.1024
[5] S. Quazi, "Artificial intelligence and machine learning in precision and genomic medicine," *Med. Oncol.*, vol. 39, no. 8, p. 120, 2022. https://doi.org/10.20944/preprints202110.0011.v1
[6] K. Sherin, A. P. A. Victoria, S. Harini, and J. C. Jensen, "Automated Diagnosis of Anemia Signs using Machine Learning," in *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 2024, pp. 1–6. https://doi.org/10.1109/ICRITO61523.2024.10522270
[7] S. Zhang and J. Song, "A chatbot based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model," *Sci. Rep.*, vol. 14, no. 1, p. 17118, 2024. https://doi.org/10.1038/s41598-024-67429-4
[8] A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, "A review on bayesian deep learning in healthcare: Applications and challenges," *IEEE Access*, vol. 10, pp. 36538–36562, 2022. https://doi.org/10.1109/ACCESS.2022.3163384
[9] B. Omarov, M. Baikuvekov, Z. Momynkulov, A. Kassenkhan, S. Nuralykyzy, and M. Iglikova, "Convolutional LSTM Network for Heart Disease Diagnosis on Electrocardiograms.," *Comput. Mater. \& Contin.*, vol. 76, no. 3, 2023.
[10] M. Saleem, W. Aslam, M. I. U. Lali, H. T. Rauf, and E. A. Nasr, "Predicting Thalassemia Using Feature Selection Techniques: A Comparative Analysis," *Diagnostics*, vol. 13, no. 22, p. 3441, 2023. https://doi.org/10.3390/diagnostics13223441
[11] J. Brooks, "Statistical Tools for Efficient Confirmation of Diagnosis in Patients with Suspected Primary Central Nervous System Vasculitis," Université d'Ottawa/University of Ottawa, 2023.
[12] R. Shouval, J. A. Fein, B. Savani, M. Mohty, and A. Nagler, "Machine learning and artificial intelligence in haematology," *Br. J. Haematol.*, vol. 192, no. 2, pp. 239–250, 2021. https://doi.org/10.1111/bjh.16915
[13] 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. https://doi.org/10.1007/s00431-020-03597-5
[14] S. Sundararajan and H. Rabe, "Prevention of iron deficiency anemia in infants and toddlers," *Pediatr. Res.*, vol. 89, no. 1, pp. 63–73, 2021. https://doi.org/10.1038/s41390-020-0907-5
[15] S. Mahmud, T. B. Donmez, M. Mansour, M. Kutlu, and C. Freeman, "Anemia detection through non-invasive analysis of lip mucosa images," *Front. big Data*, vol. 6, p. 1241899, 2023. https://doi.org/10.3389/fdata.2023.1241899
[16] K. C. Sahoo, A. Sinha, R. K. Sahoo, S. S. Suman, D. Bhattacharya, and S. Pati, "Diagnostic validation and feasibility of a non-invasive haemoglobin screening device (EzeCheck) for'Anaemia Mukt Bharat'in India," *Cureus*, vol. 16, no. 1, 2024. https://doi.org/10.7759/cureus.52877
[17] S. Bodapati, H. Bandarupally, R. N. Shaw, and A. Ghosh, "Comparison and analysis of RNN-LSTMs and CNNs for social reviews classification," *Adv. Appl. Data-Driven Comput.*, pp. 49–59, 2021. https://doi.org/10.1007/978-981-33-6919-1_4
[18] N. Rane, S. Choudhary, and J. Rane, "Ensemble Deep Learning and Machine Learning: Applications, Opportunities, Challenges, and Future Directions," *Oppor. Challenges, Futur. Dir. (May 31, 2024)*, 2024.
[19] J. A. Esterhuizen, B. R. Goldsmith, and S. Linic, "Interpretable machine learning for knowledge generation in heterogeneous catalysis," *Nat. Catal.*, vol. 5, no. 3, pp. 175–184, 2022. https://doi.org/10.1038/s41929-022-00744-z
[20] E. Shehab, A. Khawaga, and others, "Anemia Diagnosis And Prediction Based On Machine Learning," *Kafrelsheikh J. Inf. Sci.*, vol. 4, no. 2, pp. 1–9, 2023. https://doi.org/10.1007/978-981-99-0071-8_18
[21] S. Akter *et al.*, "AD-CovNet: An exploratory analysis using a hybrid deep learning model to handle data imbalance, predict fatality, and risk factors in Alzheimer's patients with COVID-19," *Comput. Biol. Med.*, vol. 146, p. 105657, 2022. https://doi.org/10.1016/j.compbiomed.2022.105657
[22] E. Aboelnaga, "Anemia Types Classification." 2023. https://www.kaggle.com/datasets/ehababoelnaga/anemia-types-classification

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 5, No. 4, August 2024, Page 1108-1117
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v5i4.5848

[23] A. Abrol *et al.*, "Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning," *Nat. Commun.*, vol. 12, no. 1, p. 353, 2021. https://doi.org/10.1038/s41467-020-20655-6
[24] A. Carè, "Gender imbalance in medical imaging datasets for Artificial Intelligence," *IGMCONGRESS 2022*, p. 30, 2022. https://doi.org/10.1073/pnas.1919012117