Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1362−1369
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6790

Performance Evaluation of Machine Learning Models for HIV/AIDS Classification

Gregorius Airlangga*

Department of Information Systems, Atma Jaya Catholic University of Indonesia, Jakarta
Jl. Jend. Sudirman No.51 5, RT.004/RW.4, Karet Semanggi, Setiabudi District, South Jakarta City, Special Capital Region of Jakarta, Indonesia
Email: gregorius.airlangga@atmajaya.ac.id
Correspondence Author Email: gregorius.airlangga@atmajaya.ac.id
Submitted: **19/01/2025;** Accepted: **31/01/2025;** Published: **31/01/2025**

Abstract−Accurate and early diagnosis of HIV/AIDS is critical for effective treatment and reducing disease transmission. This study evaluates the performance of several machine learning models, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes, for classifying HIV/AIDS infection status. A dataset comprising 50,000 samples was used, and models were assessed based on accuracy, precision, recall, and F1 score using stratified ten-fold cross-validation to ensure robust evaluation. The results reveal significant trade-offs between sensitivity and specificity across the models. Gradient Boosting achieved the highest accuracy (70.85%) and precision (57.81%), making it suitable for confirmatory testing where minimizing false positives is critical. Conversely, Naive Bayes demonstrated the highest recall (57.99%) and F1 score (51.04%), emphasizing its effectiveness in early-stage diagnostics where sensitivity is paramount. SVM exhibited the highest precision (59.87%) but struggled with recall (11.28%), reflecting its conservative nature in classifying positive cases. These findings underscore the importance of selecting models tailored to specific diagnostic objectives. While Naive Bayes is ideal for comprehensive screening programs, Gradient Boosting and SVM are better suited for confirmatory testing. This research provides valuable insights into the strengths and limitations of machine learning models for medical diagnostics, paving the way for developing more robust, hybrid approaches to optimize sensitivity and specificity in HIV/AIDS classification.

Keywords: Machine Learning; HIV/AIDS Classification; Diagnostic Models; Comparative Analysis; Evaluation Metrics

1. INTRODUCTION

Human Immunodeficiency Virus (HIV) and Acquired Immunodeficiency Syndrome (AIDS) continue to pose significant public health challenges worldwide, particularly in resource-limited settings[1]–[3]. Despite advancements in medical science and the availability of antiretroviral therapies, accurate and early diagnosis remains a cornerstone in mitigating the spread and impact of this disease [4]–[6]. Traditional diagnostic methods often rely on clinical symptoms and confirmatory laboratory tests, which, while effective, may suffer from limitations such as high costs, limited accessibility, and extended turnaround times [7]–[9]. In this context, leveraging data-driven approaches to enhance diagnostic capabilities is a burgeoning field of research with the potential to revolutionize HIV/AIDS healthcare management [10]–[12].

Machine learning has emerged as a transformative technology in medical diagnostics, offering capabilities to analyze complex datasets, identify subtle patterns, and produce robust predictions [10], [13], [14]. By automating the identification of infection statuses, machine learning algorithms can significantly augment the diagnostic process, leading to faster and more accurate decision-making [15]–[17]. The use of supervised learning algorithms for HIV/AIDS classification presents an opportunity to optimize the diagnostic pipeline, addressing challenges such as imbalanced data, feature selection, and model interpretability [18]–[20]. These advancements align with the global imperative to achieve the United Nations' Sustainable Development Goal (SDG) 3, which focuses on ensuring healthy lives and promoting well-being for all at all ages [21]–[23]. Numerous studies have explored the application of machine learning in disease diagnosis, ranging from logistic regression models for binary classification to ensemble techniques such as Random Forest and Gradient Boosting for complex decision-making scenarios [24]–[26]. However, most existing research focuses on either small datasets or lacks a comparative evaluation of models using advanced validation techniques [27]–[29]. There is a paucity of studies that examine the performance of machine learning models on large-scale datasets with stratified cross-validation, a critical requirement to ensure the generalizability and robustness of predictive models [30]–[32].

To address these gaps, this study presents a comprehensive analysis of machine learning models for the classification of HIV/AIDS infection status using a large dataset comprising 50,000 samples. The dataset encompasses a diverse set of features that reflect various clinical and demographic parameters. By employing a rigorous methodology that includes data preprocessing, feature scaling, and stratified ten-fold cross-validation, this research evaluates the performance of multiple machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and XGBoost. The contributions of this study are fourfold. First, it introduces a robust machine learning pipeline that incorporates feature scaling and stratified cross-validation to ensure reliable and unbiased model evaluation. Second, it performs a detailed comparative analysis of classical and ensemble machine learning models using multiple performance metrics, including accuracy, precision, recall, and F1 score. Third, it emphasizes the practical implications of these models in real-world diagnostic scenarios, highlighting their

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1362−1369
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6790

potential to reduce diagnostic delays and improve accessibility. Finally, this research provides actionable insights into the selection of optimal machine learning models for HIV/AIDS classification, offering a valuable reference for both academic researchers and healthcare practitioners. The remainder of this article is structured as follows. The subsequent section outlines the methodology employed in this study, including dataset description, preprocessing steps, and the evaluation framework. The results section presents the comparative performance of the machine learning models, followed by an in-depth discussion of the findings. Finally, the conclusion highlights the key contributions, limitations, and future directions of this research.

2. RESEARCH METHODOLOGY

This study develops a robust and systematic methodology to evaluate the performance of various machine learning models in classifying HIV/AIDS infection status. The methodology is structured into dataset preparation, preprocessing, model design, stratified cross-validation, performance evaluation, and implementation as presented in the figure 1.

2.1 Dataset

The dataset used in this study comprises $(N = 50{,}000)$ samples and can be downloaded from [33], each represented by (M) features capturing clinical and demographic attributes. Let the dataset be denoted as (1).

$$\mathcal{D} = \{(x_i, y_i) \mid i = 1,2, \dots, N\} \quad (1)$$

where $(x_i = [x_{i1}, x_{i2}, \dots, x_{iM}]^\top \in R^M)$ represents the feature vector for the (i)-th sample, and $(y_i \in \{0,1\})$ is the corresponding binary target variable. Here, $(y_i = 1)$ signifies an HIV-positive status, and $(y_i = 0)$ represents HIV-negative status. This supervised learning problem seeks to approximate the underlying conditional probability $(P(y \mid x))$ and produce a classifier $(f_\theta(x))$ parameterized by (θ), such that $(f_\theta(x) \approx y)$.

2.2 Data Preprocessing

Preprocessing the dataset is crucial to ensure that the models are trained on data with uniform scaling and minimal bias. Each feature (x_{ij}) in the raw dataset is normalized using Min-Max scaling, defined as (2).

$$x'_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (2)$$

where $(\min(x_j))$ and $(\max(x_j))$ represent the minimum and maximum values of the (j)-th feature across all samples. This transformation maps the feature values to the range $([0,1])$, ensuring numerical stability and preventing features with larger magnitudes from dominating the model training process. Given the class imbalance that often occurs in medical datasets, the class distributions (n_0) and (n_1), representing the number of samples in the negative and positive classes, respectively, were analyzed. The imbalance ratio, defined as (3).

$$\text{IR} = \frac{\min(n_0, n_1)}{\max(n_0, n_1)} \quad (3)$$

was computed. If $(\text{IR} < 0.7)$, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic samples for the minority class. SMOTE constructs synthetic samples by interpolating between existing minority class samples. For a minority sample (x_i), a synthetic sample $(x_{synthetic})$ is generated as (4).

$$x_{synthetic} = x_i + \delta \cdot (x_{\text{niho}} - x_i) \quad (4)$$

where (x_{niho}) is a randomly selected nearest neighbor of (x_i), and $(\delta \sim \mathcal{U}(0,1))$.

2.3 Machine Learning Models

The core of this research lies in the implementation and evaluation of various machine learning models. Logistic regression optimizes the binary cross-entropy loss, defined as (5).

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \sigma(x_i^\top \theta) + (1 - y_i)\log\left(1 - \sigma(x_i^\top \theta)\right)\right] \quad (5)$$

where $(\sigma(z) = \frac{1}{1+e^{-z}})$ is the sigmoid function. Decision trees partition the feature space into hyper-rectangles by iteratively selecting features and thresholds that minimize impurity measures such as Gini impurity, $G(t) = 1 - \sum_{c=1}^{C} p_c^2$, or entropy, $H(t) = -\sum_{c=1}^{C} p_c \log p_c$, where (p_c) is the proportion of samples in class (c) within a node (t). Furthermore, random forests extend decision trees by creating an ensemble of (T) trees, where each tree is trained on a bootstrap sample of the dataset, and features are randomly subsampled at each split. The ensemble prediction is obtained via majority voting, expressed as (6).

$$\hat{y} = \text{mode}\left(\{f_{\theta_t}(x) \mid t = 1, \dots, T\}\right) \quad (6)$$

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1362−1369
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6790

Gradient boosting improves upon weak learners iteratively by minimizing the residual errors at each stage. The optimization problem is formulated as (7).

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta L(\mathcal{D}, \theta) \quad (7)$$

where (η) is the learning rate and $(\nabla_\theta L(\mathcal{D}, \theta))$ is the gradient of the loss function (L) with respect to the model parameters. Support vector machines construct a hyperplane $(w^{\top x} + b = 0)$ that maximizes the margin between classes. The optimization problem is defined as $\min_{w,b} \quad \frac{1}{2}|w|^2$, subject to $(y_i(w^{\top x_i} + b) \geq 1)\ for\ all(i)$.

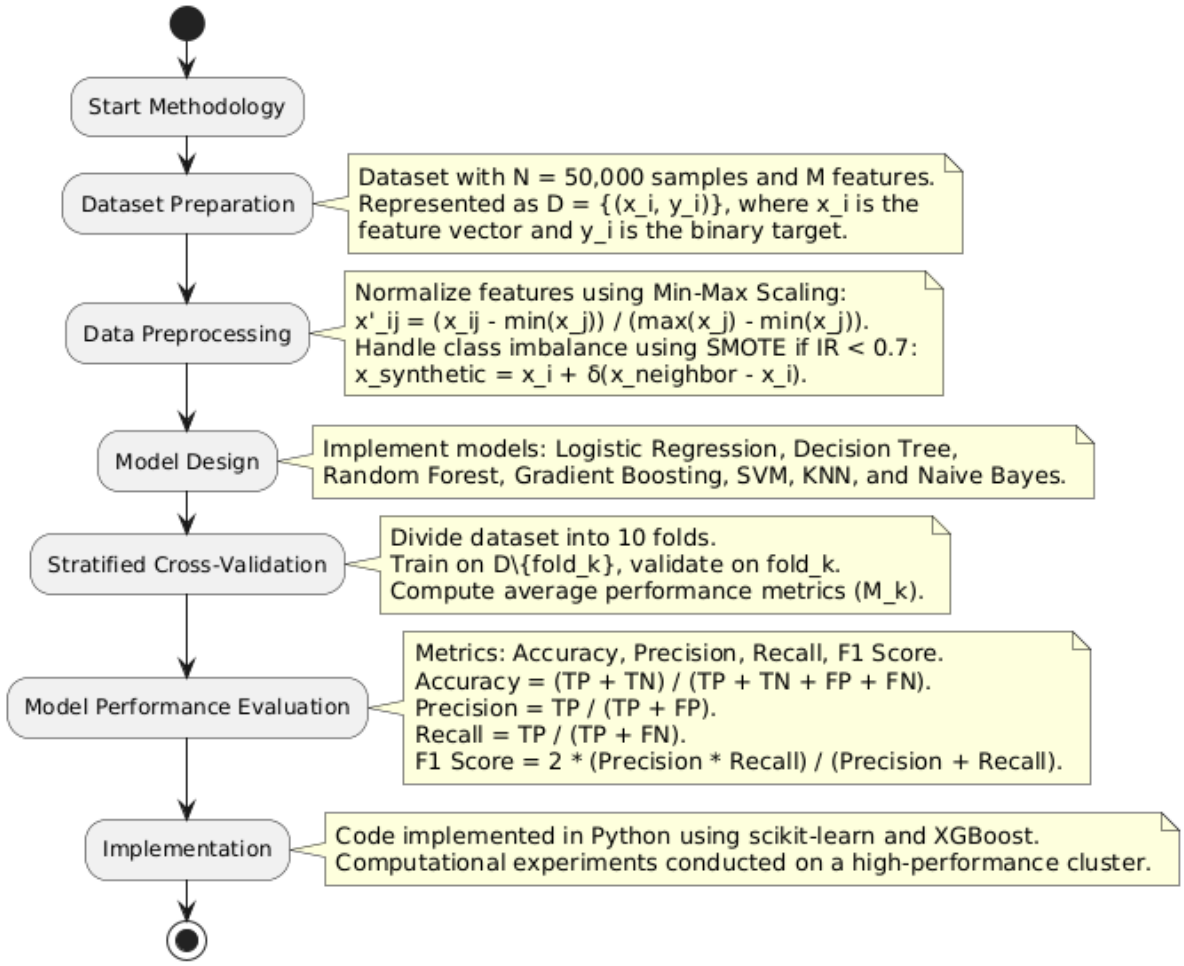

Figure 1. Research Methodology

The optimization is solved using Lagrange multipliers and the kernel trick for non-linear decision boundaries. In addition, K-nearest neighbors (KNN) predicts the class of a sample (x) by majority voting among its (k) nearest neighbors, where the distance metric $(d(x, x') = |x - x'|_2)$ is used. Naive Bayes computes posterior probabilities using Bayes' theorem under the assumption of feature independence as (8).

$$P(y \mid \boldsymbol{x}) \propto P(y) \prod_{j=1}^{M} P(x_j \mid y) \quad (8)$$

2.4 Evaluation

To evaluate model performance, stratified ten-fold cross-validation was employed. The dataset was divided into ten equally sized folds, $(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{10})$, ensuring that the class distribution within each fold mirrors that of the entire dataset. For each fold (k), the model was trained on the training set $(\mathcal{D} \setminus \mathcal{D}_k)$ and tested on the validation set $(\mathcal{D}_k)$. The performance metric (M) was computed as the average across folds as presented as (9).

$$M = \frac{1}{10} \sum_{k=1}^{10} M_k \quad (9)$$

where (M_k) is the metric value for fold (k). Performance metrics included accuracy, precision, recall, and F1 score. Accuracy is computed as (10).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (10)$$

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1362−1369
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6790

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Precision is calculated as (11).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{11}$$

while recall is presented in (12).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{12}$$

The F1 score, a harmonic mean of precision and recall, is defined as (13)

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \tag{13}$$

All computational experiments were conducted in Python, utilizing the scikit-learn and XGBoost libraries. The experiments were executed on a high-performance computing cluster to ensure scalability and computational efficiency. This rigorous methodology ensures the robustness and generalizability of the evaluated models, providing a strong foundation for analyzing machine learning performance in HIV/AIDS classification tasks.

3. RESULT AND DISCUSSION

This section presents a comprehensive analysis of the performance of several machine learning models applied to the classification of HIV/AIDS infection status. Each model was evaluated using stratified ten-fold cross-validation to ensure a robust and unbiased assessment. The evaluation metrics as presented in the Table 1, include accuracy, precision, recall, and F1 score, with both the mean and standard deviation reported for each metric. These metrics are critical for understanding the predictive reliability, sensitivity, and balance between false positives and false negatives for each model, especially in the context of a medical application where misclassification can have significant consequences.

3.1 Result

The logistic regression model achieved a mean accuracy of 0.7067 with a standard deviation of 0.0028, demonstrating stable and consistent performance across folds. However, its recall value of 0.2269 indicates a significant limitation in detecting positive cases. This low recall suggests that the model fails to identify a substantial proportion of true positives, which is particularly concerning in a diagnostic setting where sensitivity is paramount. The precision of 0.5682 and F1 score of 0.3242 further reflect the model's inability to balance precision and recall effectively. While logistic regression provides a relatively high overall accuracy, its inability to capture positive cases makes it unsuitable for use as a standalone diagnostic tool in this context.

Table 1. Performance Results

Model	Accuracy	F1 Score	Precision	Recall
Decision Tree	0.61558	0.38972	0.38170	0.40371
Gradient Boosting	0.70850	0.32196	0.57813	0.22320
KNN	0.66244	0.36632	0.43844	0.31465
Logistic Regression	0.70672	0.32418	0.56819	0.22688
Naive Bayes	0.65492	0.51041	0.45584	0.57997
Random Forest	0.70532	0.33261	0.55912	0.23617
SVM	0.70132	0.18968	0.59875	0.11280

The decision tree model, with an accuracy of 0.6156, exhibited a trade-off between precision (0.3817) and recall (0.4037). The F1 score of 0.3897 underscores a marginal improvement in balancing sensitivity and precision compared to logistic regression. Decision trees are known for their interpretability, but their susceptibility to overfitting likely contributed to the moderate standard deviation in accuracy (0.0086). This variability suggests that the model's performance may be inconsistent when exposed to unseen data, limiting its generalizability in practical applications. The random forest model, an ensemble-based approach, achieved an accuracy of 0.7053, comparable to logistic regression, but with slightly higher variability as indicated by a standard deviation of 0.0041. Its recall of 0.2362 and precision of 0.5591 resulted in an F1 score of 0.3326, which, while slightly better than logistic regression, still reflects an imbalance in sensitivity and specificity. Random forests excel in capturing complex feature interactions and tend to generalize well on large datasets. However, the marginal improvement over logistic regression suggests that the dataset's characteristics may not strongly favor ensemble techniques, likely due to limited feature complexity or interdependencies.

Gradient boosting emerged as the most accurate model with a mean accuracy of 0.7085 and the highest precision of 0.5781 among all models evaluated. Its low recall value of 0.2232, however, limits its effectiveness in identifying positive cases, as reflected in the F1 score of 0.3220. Gradient boosting's strength lies in its iterative refinement of weak learners, which enables it to identify subtle patterns in the data. The high precision suggests

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1362−1369
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6790

that the model is highly reliable in predicting positive cases but at the cost of a significant number of false negatives. This trade-off may render gradient boosting less suitable for a diagnostic application where recall is critical. The support vector machine (SVM) model achieved a mean accuracy of 0.7013 with a very low standard deviation of 0.0013, indicating exceptional stability across cross-validation folds. Despite this stability, the recall of 0.1128 was the lowest among all models, highlighting a severe deficiency in identifying positive cases. With a precision of 0.5987, the SVM model appears highly conservative, favoring correct positive predictions over broad sensitivity. The resulting F1 score of 0.1897, the lowest across all models, underscores the model's inadequacy in balancing precision and recall. This outcome aligns with the inherent characteristics of SVM, which can prioritize margin maximization over handling imbalanced data effectively.

The k-nearest neighbors (KNN) model exhibited a moderate accuracy of 0.6624 and a recall of 0.3147, reflecting a more balanced sensitivity compared to other models like SVM or logistic regression. Its precision of 0.4384 resulted in an F1 score of 0.3663, which, while not the highest, suggests a relatively better trade-off between precision and recall. KNN's performance depends heavily on the choice of the number of neighbors (k) and the underlying distance metric, which may explain its moderate performance relative to more sophisticated models like gradient boosting or random forest. Naive Bayes, despite its simplistic assumption of feature independence, demonstrated unique strengths. It achieved the lowest accuracy of 0.6549 but excelled in recall with a value of 0.5800, the highest among all models. Its precision of 0.4558 and F1 score of 0.5104 further highlight its capability to identify positive cases effectively. These results suggest that Naive Bayes prioritizes sensitivity, making it well-suited for scenarios where false negatives must be minimized, such as early-stage disease detection. However, its reliance on the independence assumption likely limits its ability to capture complex feature interactions, contributing to its lower accuracy compared to ensemble models.

3.2 Discussion

The comparative analysis of these models highlights important trade-offs that must be considered in the context of HIV/AIDS classification. Gradient boosting achieved the highest accuracy and precision, making it suitable for applications where overall predictive reliability and minimizing false positives are prioritized. Conversely, Naive Bayes, with its high recall and F1 score, is more appropriate for scenarios emphasizing sensitivity, where missing positive cases can have severe consequences. These results suggest that a single model may not suffice for all diagnostic objectives. Instead, a hybrid approach that combines the strengths of precision-focused and recall-focused models may be a promising direction for future research. The limitations observed across models also emphasize the need for further refinements, such as feature engineering to enhance discriminatory power, hyperparameter optimization to fine-tune model performance, and the incorporation of advanced techniques like ensemble stacking to improve overall robustness. Additionally, exploring cost-sensitive learning frameworks or leveraging domain-specific knowledge could further mitigate the trade-offs between sensitivity and specificity observed in this study.

3.3 Statistical Analysis of Metric Variability

The standard deviations of the evaluation metrics provide valuable insights into the consistency and reliability of the models across cross-validation folds. Logistic Regression exhibited the lowest variability in accuracy (0.0028), indicating highly stable performance across all folds. This consistency is particularly advantageous for practical deployment, as it minimizes the risk of unexpected deviations in performance on new data. Similarly, Gradient Boosting and Random Forest demonstrated relatively low standard deviations in accuracy (0.0049 and 0.0041, respectively), reinforcing their reliability as ensemble methods. Naive Bayes achieved the lowest standard deviation in recall (0.0062), showcasing its robustness in detecting positive cases consistently. This stability is crucial in medical applications where the sensitivity of a model can directly impact patient outcomes. The model's probabilistic foundation likely contributes to its consistent recall performance, even when trained on varying subsets of the data.

Conversely, SVM displayed the lowest standard deviation in accuracy (0.0013), signifying its high reliability in overall classification. However, its high variability in recall (0.0066) suggests inconsistency in detecting positive cases, which may undermine its applicability in scenarios requiring high sensitivity. This instability in recall is consistent with SVM's focus on margin maximization, which can lead to variability in handling minority classes in imbalanced datasets. Gradient Boosting, despite its strong accuracy and precision, exhibited moderate variability in recall (0.0090). This variability underscores the need for careful hyperparameter tuning to ensure consistent performance. Random Forest, with a recall standard deviation of 0.0080, demonstrated slightly better stability than Gradient Boosting but still required optimization for consistent sensitivity.

3.4 Computational Complexity of Machine Learning Models

Understanding the computational time and space complexity of each machine learning model is crucial for evaluating their practicality, especially when working with large-scale datasets like the one used in this study. Logistic Regression has a time complexity of $O(n \cdot p \cdot k)$, where n represents the number of samples, p the number

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1362−1369
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6790

of features, and k the number of iterations required for convergence, which depends on the optimization algorithm employed. Its space complexity is relatively low, at $O(p)$, as it only needs to store the coefficients of the features.

Decision Tree models have a time complexity of $O(n \cdot p \cdot \log n)$, which is primarily determined by the need to sort data for each feature during the split process. The space complexity for Decision Trees is $O(n \cdot p)$, as the model must store the structure of the tree and the data at each node. Random Forest models extend the complexity of Decision Trees by considering an ensemble of m trees. Their time complexity increases to $O(m \cdot n \cdot p \cdot \log n)$, while their space complexity scales to $O(m \cdot n)$ due to the storage requirements for multiple trees. Gradient Boosting shares similarities with Random Forest in terms of its ensemble nature but focuses on iterative refinement of weak learners. Its time complexity is $O(m \cdot n \cdot \log n)$, where m denotes the number of boosting iterations, and its space complexity remains $O(m \cdot n)$. Support Vector Machines (SVMs) exhibit varying computational complexities depending on the kernel used. For linear kernels, the time complexity is $O(n \cdot p)$, while for non-linear kernels such as radial basis function (RBF), it can increase to $O(n^2 \cdot p)$ or $O(n^3)$ during training. The space complexity of SVMs is $O(n^2)$, primarily due to the storage of kernel matrices.

K-Nearest Neighbors (KNN) models are computationally expensive during prediction, with a time complexity of $O(n \cdot p \cdot k)$, where k is the number of nearest neighbors considered. The space complexity of KNN is $O(n \cdot p)$ because it stores all the training samples. Naive Bayes, on the other hand, is computationally efficient, with a time complexity of $O(n \cdot p)$. Its space complexity is $O(p \cdot c)$, where c represents the number of classes, as it assumes feature independence and only requires the storage of probabilities for each class-feature combination. Finally, XGBoost, a popular ensemble method, has a time complexity of $O(m \cdot n \cdot \log n)$, where m is the number of boosting iterations. Its space complexity is $O(m \cdot n)$, comparable to that of Gradient Boosting. These computational considerations provide a comprehensive understanding of the trade-offs between the models in terms of their resource demands, guiding their selection based on the specific requirements of the application.

3.5 Trade-Offs Between Metrics and Practical Implications

The results of this study reveal significant trade-offs between precision and recall across the evaluated models, which have critical implications for their practical deployment in HIV/AIDS diagnostics. Precision and recall are two complementary metrics that often present conflicting objectives in machine learning, especially in medical applications where the costs of false positives and false negatives differ greatly. Models such as Support Vector Machine (SVM) and Gradient Boosting are designed to prioritize precision, achieving values of 0.5987 and 0.5781, respectively. These high precision scores indicate the models' strong ability to correctly identify positive cases with minimal false positives. This behavior makes them particularly well-suited for confirmatory testing, where minimizing false positives is paramount to reduce unnecessary anxiety for patients and avoid wasting medical resources on follow-up procedures for misclassified cases.

However, the emphasis on precision comes at the expense of recall, as observed in SVM and Gradient Boosting, which achieved recall values of only 0.1128 and 0.2232, respectively. Such low recall indicates that these models fail to identify a substantial portion of true positive cases, which can be detrimental in scenarios where missing positive diagnoses poses significant risks. For instance, undetected cases of HIV/AIDS can lead to delayed treatment, increased disease transmission, and worsening public health outcomes. This trade-off highlights the inherent limitations of precision-focused models when applied to early-stage diagnostic applications, where the primary objective is to capture as many positive cases as possible.

In contrast, Naive Bayes takes the opposite approach by emphasizing recall, achieving the highest recall score of 0.5800 among all models. This high recall ensures that the majority of positive cases are identified, making Naive Bayes particularly advantageous for early screening programs. Early-stage diagnosis often prioritizes sensitivity over specificity, as the consequences of false negatives: undetected positive cases are far more severe than false positives. However, Naive Bayes achieves this high recall at the expense of precision (0.4558), which translates to a higher number of false positives. While this could result in increased follow-up testing, the model's ability to minimize false negatives aligns well with the objectives of public health initiatives aimed at early detection and intervention.

The F1 score, which balances precision and recall, provides a holistic perspective on model performance. Naive Bayes achieved the highest F1 score of 0.5104, demonstrating its effectiveness in maintaining a reasonable balance between sensitivity and specificity. This metric highlights the model's suitability for scenarios where both metrics are critical, such as large-scale screening programs where resources must be allocated efficiently while ensuring comprehensive case identification. Gradient Boosting and Random Forest, while achieving high precision, exhibited lower F1 scores (0.3220 and 0.3326, respectively), reflecting their limited sensitivity in detecting positive cases despite their accuracy and specificity.

From a practical standpoint, the trade-offs between precision and recall must be carefully considered based on the specific objectives of the diagnostic application. For early-stage diagnosis, where the goal is to identify all potential positive cases and minimize the risk of false negatives, Naive Bayes emerges as the most appropriate choice due to its high recall and balanced F1 score. Its robustness in capturing true positives ensures that fewer cases go undetected, which is essential for initiating timely treatment and mitigating disease spread. While the

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1362−1369
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6790

higher false positive rate associated with Naive Bayes may increase the burden of follow-up testing, this trade-off is acceptable in scenarios where the stakes of missing positive cases are high.

Conversely, for confirmatory testing, where the focus shifts to reducing false positives and ensuring that identified cases are indeed true positives, models like Gradient Boosting and SVM may be more suitable. Their high precision ensures reliability in positive classifications, reducing the likelihood of misdiagnosis and unnecessary interventions. This makes them valuable tools for refining diagnostic pipelines, where initial screening is followed by more rigorous confirmatory testing to validate cases.

4. CONCLUSION

This study evaluated the performance of several machine learning models for HIV/AIDS classification using a dataset of 50,000 samples. The analysis focused on accuracy, precision, recall, and F1 score, highlighting significant trade-offs between these metrics. Gradient Boosting demonstrated the highest accuracy (0.7085) and precision (0.5781), making it suitable for confirmatory testing where reducing false positives is critical. In contrast, Naive Bayes exhibited the highest recall (0.5800) and F1 score (0.5104), emphasizing its effectiveness in early-stage diagnostics where sensitivity is paramount to avoid missed cases. Models such as Logistic Regression and Random Forest showed strong accuracy but struggled with recall, while SVM achieved high precision (0.5987) but had the lowest recall (0.1128). The findings underscore the need for tailored model selection based on diagnostic objectives. Naive Bayes is well-suited for comprehensive screening programs, while Gradient Boosting and SVM are more appropriate for confirmatory diagnostics. Future research should explore hybrid approaches combining the strengths of different models to enhance both sensitivity and specificity. Additionally, advanced techniques like feature engineering and domain-specific cost-sensitive learning could further improve model performance. This study provides valuable insights for developing robust and clinically relevant machine learning models, contributing to improved HIV/AIDS diagnosis and public health outcomes.

ACKNOWLEDGMENT

The authors would like to express their heartfelt gratitude to the Atma Jaya Catholic University of Indonesia for its invaluable support and resources that made this research possible. The university's dedication to fostering academic excellence and providing an environment conducive to innovation and discovery has been instrumental in the successful completion of this study. We also extend our appreciation to the Faculty of Engineering and the Department of Information Systems for their encouragement and guidance throughout this research journey. Their commitment to advancing knowledge in the fields of machine learning and healthcare has greatly inspired this work.

REFERENCES

[1] M. Mark, “The international problem of HIV/AIDS in the modern world: a comprehensive review of political, economic, and social impacts,” *Res Output J Public Heal. Med*, vol. 42, pp. 47–52, 2024.
[2] K. Balogun and P. R. Slev, “Towards achieving the end of the HIV epidemic: advances, challenges and scaling-up strategies,” *Clin. Biochem.*, vol. 117, pp. 53–59, 2023.
[3] B. Parekh, L. Westerman, L. Vojnov, and C. Yang, “Human immunodeficiency virus (HIV) infections,” *Lab. point-of-care diagnostic Test. Sex. Transm. Infect. Incl. HIV*, p. 247, 2023.
[4] J. Prajapati, A. Kumari, and others, “Navigating the Spectrum: A Comprehensive Review of HIV Detection Methods,” 2024.
[5] K. Lakshmanan and B. M. Liu, “Impact of Point-of-Care Testing on Diagnosis, Treatment, and Surveillance of Vaccine-Preventable Viral Infections,” *Diagnostics*, vol. 15, no. 2, p. 123, 2025.
[6] E. I. Obeagu and G. U. Obeagu, “Early Infant Diagnosis: A Crucial Step in Halting HIV Transmission,” *Elit. J. Heal. Sci.*, vol. 1, no. 1, pp. 1–11, 2023.
[7] S. Chakraborty, “Democratizing nucleic acid-based molecular diagnostic tests for infectious diseases at resource-limited settings--from point of care to extreme point of care,” *Sensors \& Diagnostics*, vol. 3, no. 4, pp. 536–561, 2024.
[8] K. Patel *et al.*, “Forty years since the epidemic: modern paradigms in HIV diagnosis and treatment,” *Cureus*, vol. 13, no. 5, 2021.
[9] A. Bacon *et al.*, “Review of HIV self testing technologies and promising approaches for the next generation,” *Biosensors*, vol. 13, no. 2, p. 298, 2023.
[10] A. Padhi, A. Agarwal, S. K. Saxena, and C. D. S. Katoch, “Transforming clinical virology with AI, machine learning and deep learning: a comprehensive review and outlook,” *VirusDisease*, vol. 34, no. 3, pp. 345–355, 2023.
[11] O. D. Balogun *et al.*, “Integrating ai into health informatics for enhanced public health in Africa: a comprehensive review,” *Int. Med. Sci. Res. J.*, vol. 3, no. 3, pp. 127–144, 2023.
[12] A. M. Zaidan, “The leading global health challenges in the artificial intelligence era,” *Front. Public Heal.*, vol. 11, p. 1328918, 2023.
[13] A. Sharma, A. Lysenko, S. Jia, K. A. Boroevich, and T. Tsunoda, “Advances in AI and machine learning for predictive medicine,” *J. Hum. Genet.*, pp. 1–11, 2024.
[14] K. P. Reddy, M. Satish, A. Prakash, S. M. Babu, P. P. Kumar, and B. S. Devi, “Machine Learning Revolution in Early

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1362−1369
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6790

Disease Detection for Healthcare: Advancements, Challenges, and Future Prospects," in *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, 2023, pp. 638–643.
[15] S. Asif *et al.*, "Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision," *Arch. Comput. Methods Eng.*, pp. 1–31, 2024.
[16] M. M. Siddique, M. M. Bin Seraj, M. N. Adnan, and S. M. Galib, "Artificial Intelligence for Infectious Disease Detection: Prospects and Challenges," *Surveillance, Prev. Control Infect. Dis. An AI Perspect.*, pp. 1–22, 2024.
[17] N. Hong *et al.*, "State of the art of machine learning--enabled clinical decision support in intensive care units: literature review," *JMIR Med. informatics*, vol. 10, no. 3, p. e28781, 2022.
[18] G. Obaido *et al.*, "An interpretable machine learning approach for hepatitis b diagnosis," *Appl. Sci.*, vol. 12, no. 21, p. 11127, 2022.
[19] H. H. Rashidi, L. T. Dang, S. Albahra, R. Ravindran, and I. H. Khan, "Automated machine learning for endemic active tuberculosis prediction from multiplex serological data," *Sci. Rep.*, vol. 11, no. 1, p. 17900, 2021.
[20] K. I. Sahibzada *et al.*, "HIV OctaScanner: A Machine Learning Approach to Unveil Proteolytic Cleavage Dynamics in HIV-1 Protease Substrates," *J. Chem. Inf. Model.*, 2025.
[21] S. Sorooshian, "The Sustainable Development Goals of the United Nations: A Comparative Midterm Research Review," *J. Clean. Prod.*, p. 142272, 2024.
[22] B. Y. F. Fong, V. T. S. Law, T. C. H. Leung, M. F. Lo, T. K. C. Ng, and H. H. L. Yee, *Sustainable development goal 3: Health and well-being of ageing in Hong Kong*. Taylor \& Francis, 2022.
[23] G. Venkatesh, "A brief analysis of SDG 3--Good health and well-being--and its synergies and trade-offs with the other Sustainable Development Goals," *Probl. Ekorozwoju*, vol. 17, no. 2, 2022.
[24] 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.
[25] V. R. Modhugu and S. Ponnusamy, "Comparative Analysis of Machine Learning Algorithms for Liver Disease Prediction: SVM, Logistic Regression, and Decision Tree," *Asian J. Res. Comput. Sci.*, vol. 17, no. 6, pp. 188–201, 2024.
[26] Z. Rahmatinejad *et al.*, "A comparative study of explainable ensemble learning and logistic regression for predicting in-hospital mortality in the emergency department," *Sci. Rep.*, vol. 14, no. 1, p. 3406, 2024.
[27] J. Pachouly, S. Ahirrao, K. Kotecha, G. Selvachandran, and A. Abraham, "A systematic literature review on software defect prediction using artificial intelligence: Datasets, Data Validation Methods, Approaches, and Tools," *Eng. Appl. Artif. Intell.*, vol. 111, p. 104773, 2022.
[28] A. Khraisat and A. Alazab, "A critical review of intrusion detection systems in the internet of things: techniques, deployment strategy, validation strategy, attacks, public datasets and challenges," *Cybersecurity*, vol. 4, pp. 1–27, 2021.
[29] K. Li *et al.*, "Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations," *PLoS One*, vol. 16, no. 8, p. e0255809, 2021.
[30] T. Kyriazos and M. Poga, "Application of machine learning models in social sciences: managing nonlinear relationships," *Encyclopedia*, vol. 4, no. 4, pp. 1790–1805, 2024.
[31] B. Koçak, "Key concepts, common pitfalls, and best practices in artificial intelligence and machine learning: focus on radiomics," *Diagnostic Interv. Radiol.*, vol. 28, no. 5, p. 450, 2022.
[32] C. Aliferis and G. Simon, "Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI," in *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*, Springer, 2024, pp. 477–524.
[33] A. Velu, "AIDS Virus Infection Prediction." Kaggle, 2023. {https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction