Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1310−1318
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6783

Evaluating Deep Learning Models for HIV/AIDS Classification: A Comparative Study Using Clinical and Laboratory Data

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: **18/01/2025;** Accepted: **31/01/2025;** Published: **31/01/2025**

Abstract−The accurate classification of HIV/AIDS status is critical for effective diagnosis, treatment planning, and disease management. This study evaluates the performance of four deep learning models: Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) on a comprehensive clinical and laboratory dataset derived from the AIDS Clinical Trials Group Study 175. The dataset includes features such as demographic information, treatment history, and immune markers like CD4 and CD8 counts. To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied, followed by stratified 10-fold cross-validation to ensure robust evaluation. Each model's performance was assessed using metrics including accuracy, precision, recall, F1-score, and ROC-AUC. GRU emerged as the most effective model, achieving the highest accuracy (71.04%) and ROC-AUC (57.72%), demonstrating its robustness in handling sequential data. CNN and LSTM showed competitive performance, particularly in balancing precision and recall. However, all models faced challenges in recall, highlighting difficulties in identifying minority-class samples. The findings underscore the potential of GRU for HIV/AIDS classification while identifying limitations in current approaches to handling class imbalance. Future work will explore advanced architectures, such as attention mechanisms and hybrid models, to further improve sensitivity and robustness. This study contributes to the growing body of research on applying deep learning to healthcare, with implications for improving diagnostic accuracy and patient outcomes.

Keywords: HIV/AIDS Classification; Deep Learning; Clinical Data Analysis; Gated Recurrent Unit (GRU); Healthcare Diagnostics

1. INTRODUCTION

Human Immunodeficiency Virus (HIV) and Acquired Immunodeficiency Syndrome (AIDS) have posed significant global health challenges for decades [1]–[3]. Despite advancements in antiretroviral therapy (ART) and preventive measures, HIV/AIDS remains a critical issue, particularly in regions with limited healthcare infrastructure and resources [4]–[6]. The accurate and timely classification of HIV/AIDS status based on patient data is essential for effective treatment, disease management, and improving patient outcomes [7]–[9]. As medical datasets grow increasingly complex, traditional diagnostic methods often fail to leverage the vast amount of clinical data available [10]–[12]. In this context, machine learning (ML) and deep learning (DL) models offer unparalleled potential for predictive modeling and classification in healthcare [11]. The dataset used in this study originates from the AIDS Clinical Trials Group Study 175, one of the most extensive datasets containing clinical and laboratory data of patients diagnosed with AIDS [13]. The dataset includes a diverse set of attributes encompassing patient demographics, treatment history, and laboratory results such as CD4 and CD8 counts. These variables represent crucial indicators of immune system health and treatment efficacy. Given the inherent complexity and heterogeneity of this dataset, advanced ML and DL approaches are required to harness its predictive potential effectively.

Recent literature highlights the growing adoption of ML and DL techniques in healthcare for disease prediction, patient stratification, and treatment optimization [14], [15]. Methods such as Support Vector Machines (SVMs), Random Forests, and ensemble learning have demonstrated significant promise in handling structured clinical data [11], [16], [17]. However, while these models perform well on structured datasets, they often require feature engineering and are limited in capturing hierarchical patterns in data. On the other hand, DL architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs), have revolutionized data analysis by learning complex patterns directly from raw inputs [18]–[20]. In HIV/AIDS classification tasks, the adoption of DL models remains underexplored, despite their potential to address challenges posed by class imbalances, noisy data, and intricate feature interdependencies [21]–[23]. This study seeks to fill the gap in existing research by evaluating the performance of multiple DL models: Multilayer Perceptron (MLP), CNN, LSTM, and GRU on the HIV/AIDS dataset. By incorporating a robust ten-fold stratified cross-validation approach, the study ensures that the evaluation metrics are unbiased and generalizable. The models are further optimized using early stopping techniques to prevent overfitting, ensuring reliable performance across multiple folds. Additionally, preprocessing techniques, including Min-Max scaling and label encoding, are applied to prepare the dataset for DL architectures.

A unique contribution of this research lies in comparing the strengths and weaknesses of each DL model in terms of key performance metrics, including accuracy, precision, recall, F1 score, and ROC-AUC. Such comparisons not only shed light on the suitability of DL methods for clinical datasets but also provide actionable

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1310−1318
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6783

insights for practitioners aiming to deploy these techniques in real-world scenarios. The urgency of this research is underscored by the critical role that accurate HIV/AIDS classification plays in resource-limited settings. Misclassification can result in delayed or inappropriate treatments, exacerbating patient morbidity and mortality. Furthermore, as the global healthcare sector embraces digitalization, leveraging advanced computational techniques for disease classification is no longer optional but imperative. This paper is structured as follows. The next section provides an overview of related works, emphasizing the application of ML and DL techniques to healthcare data. Following this, the methodology is described in detail, covering data preprocessing, model architectures, training procedures, and evaluation metrics. The results and discussion section presents a comparative analysis of model performance and highlights the implications of these findings for HIV/AIDS classification. Finally, the conclusion summarizes the study's contributions, outlines its limitations, and suggests directions for future research.

2. RESEARCH METHODOLOGY

This section describes the methodological framework employed in this study to classify HIV/AIDS status using deep learning (DL) models. The methodology includes data preprocessing, the design of DL model architectures, training procedures, and evaluation metrics. Each step is detailed to ensure clarity and reproducibility, with mathematical rigor incorporated into the explanations. The dataset used in this study was obtained from the AIDS Clinical Trials Group Study 175. It includes a comprehensive set of attributes describing patient demographics, medical history, treatment details, and laboratory results. The target variable, denoted as y, indicates whether a patient is infected with HIV/AIDS ($y \in \{0, 1\}$, where 0 represents not infected and 1 represents infected). The features, represented as $X = \{x_1, x_2, \ldots, x_n\}$, include continuous and categorical variables. The goal is to develop DL models that predict y$ given $Ƌb{X}, optimizing classification performance across multiple evaluation metrics.

2.1 Data Preprocessing

The raw dataset required extensive preprocessing to prepare it for model training. Missing values in the dataset were addressed by applying imputation techniques. Continuous variables such as baseline CD4 counts (x_{cd40}) and CD8 counts (x_{cd80}) were imputed using their mean values, mathematically expressed as (1).

$$x_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \quad (1)$$

where x_{ij} represents the j-th value of the i-th feature. For categorical variables such as gender (x_{gender}) and race (x_{race}), mode imputation was applied, defined as (2).

$$x_i = \text{Mode}(x) = \arg\max_{v \in V} f(v) \quad (2)$$

where $f(v)$ represents the frequency of value v in feature x_i. Categorical variables were converted into numerical representations using label encoding. For a categorical variable x with categories $C = \{c_1, c_2, \ldots, c_m\}$, each category c_k was mapped to a unique integer k such that $x \in \{1, 2, \ldots, m\}$. The continuous features were scaled to the range [0,1] using Min-Max Scaling, defined mathematically as $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$, where x represents the original value, $\min(x)$ is the minimum value in the feature, and $\max(x)$ is the maximum value. Class imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE). Given the minority class samples S_m and the majority class samples S_M, SMOTE generates synthetic samples for S_m by interpolating between randomly selected pairs of samples in S_m. For two samples $s_1, s_2 \in S_m$, a synthetic sample s_{new} is generated as $s_{new} = s_1 + \lambda(s_2 - s_1)$, where λ is a random value in [0,1]. The dataset was split into features (X) and target labels (y), followed by stratified 10-fold cross-validation. In each fold, the data was divided into training (X_{train}, y_{train}) and test (X_{test}, y_{test}) subsets, maintaining the class distribution across folds.

2.2 Model Architectures

Four deep learning models were developed: Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). The architectures were designed to process inputs $X \in R^{n \times d}$, where n represents the number of samples and d represents the feature dimensions. The MLP model consists of fully connected layers. The input layer maps the features X to a high-dimensional space, represented as (3).

$$h_1 = \sigma(W_1 X + b_1) \quad (3)$$

where W_1 and b_1 are the weights and biases of the first layer, and σ is the ReLU activation function as presented as (4).

$$\sigma(z) = \max(0, z) \quad (4)$$

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1310−1318
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6783

Subsequent layers apply similar transformations, with dropout regularization applied to reduce overfitting. The output layer uses the softmax function as presented as (5).

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad (5)$$

where $\hat{y}_i$ is the predicted probability for class i, and k is the number of classes. The CNN model includes a convolutional layer with filters of size 3×1, extracting spatial patterns from the input X. The convolution operation is expressed as (6).

$$h_{conv} = \sigma(W_{conv} * X + b_{conv}) \quad (6)$$

where $*$ denotes the convolution operation. Flattening and dense layers follow to produce class probabilities. The LSTM model processes sequences X using recurrent connections to capture temporal dependencies. The hidden state h_t at time t is updated as (7).

$$h_t = \sigma(W_h X_t + U_h h_{t-1} + b_h) \quad (7)$$

Where W_h , and b_h are learnable parameters, and X_t is the input at time t. The GRU model follows a similar structure, with update and reset gates controlling the information flow.

2.3 Training Procedures

Each model was trained using the categorical cross-entropy loss function $L = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \log(\widehat{y_{ij}})$, where y_{ij} is the true label for sample i and class j, and $\widehat{y_{ij}}$ is the predicted probability. The training process used the Adam optimizer, which updates parameters as expressed as (8).

$$\theta_{t+1} = \theta_t - \eta \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}}+\epsilon} \quad (8)$$

where η is the learning rate, $\widehat{m_t}$ and $\widehat{v_t}$ are the bias-corrected first and second moments of gradients, and ϵ is a small constant. Early stopping was employed to terminate training when validation loss did not improve for 5 consecutive epochs. This reduces the risk of overfitting and ensures efficient training.

2.4 Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Accuracy is defined as (9).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. Precision and recall are given by expression (10).

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN} \quad (10)$$

The F1-score, the harmonic mean of precision and recall, is calculated as (11).

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} \quad (11)$$

ROC-AUC quantifies the trade-off between sensitivity and specificity, defined as the area under the receiver operating characteristic curve. This methodological framework ensures that the analysis is rigorous, reproducible, and provides valuable insights into the applicability of DL models for HIV/AIDS classification.

3. RESULT AND DISCUSSION

This section provides a comprehensive analysis of the results obtained from evaluating deep learning (DL) models: Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) on the task of classifying HIV/AIDS status. Each model's performance is examined in terms of accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide insight into each model's ability to make predictions, handle imbalanced data, and identify relevant patterns in the dataset. Additionally, this discussion contextualizes the findings within the broader scope of DL applications in healthcare, highlighting potential improvements and implications.

3.1 Overview of Model Performance

The results across evaluation metrics are summarized in Table 1. GRU achieved the highest accuracy (0.7104 ± 0.0052) and precision (0.5908 ± 0.0302), making it the most effective model overall. CNN and LSTM followed closely, demonstrating competitive performance across most metrics, whereas MLP consistently underperformed

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1310−1318
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6783

in comparison. Despite these differences, all models showed challenges in recall, which indicates difficulty in identifying minority-class samples (i.e., individuals diagnosed with HIV/AIDS). This imbalance affects the models' ability to achieve a strong balance between precision and recall, as reflected in the moderate F1-scores.

3.2 Accuracy Result

Accuracy measures the proportion of correctly predicted samples over the total dataset size: $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$, where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. GRU achieved the highest accuracy (0.7104 ± 0.0052), followed closely by CNN (0.7095 ± 0.0042) and LSTM (0.7094 ± 0.0044). The MLP model recorded slightly lower accuracy (0.7082 ± 0.0035). These results indicate that all models performed comparably in their overall ability to classify samples correctly. However, the marginal differences highlight the advantage of recurrent architecture (GRU and LSTM) in capturing sequential dependencies within the dataset, which likely contributed to their higher accuracy. Accuracy is presented in the Table 1 and Figure 1.

Table 1. Performance Results

Model	Metric	Mean	Std Dev
MLP	Accuracy	0.70822	0.003503
	Precision	0.580338	0.017518
	Recall	0.216882	0.024643
	F1 Score	0.314718	0.02477
	ROC-AUC	0.572986	0.007558
CNN	Accuracy	0.70952	0.004156
	Precision	0.585555	0.025673
	Recall	0.227524	0.043666
	F1 Score	0.324394	0.04127
	ROC-AUC	0.576857	0.012064
LSTM	Accuracy	0.70936	0.00443
	Precision	0.582399	0.021505
	Recall	0.227394	0.032953
	F1 Score	0.325293	0.031784
	ROC-AUC	0.576705	0.009727
GRU	Accuracy	0.71042	0.005205
	Precision	0.590787	0.030213
	Recall	0.226298	0.038007
	F1 Score	0.324505	0.035045
	ROC-AUC	0.577171	0.010183

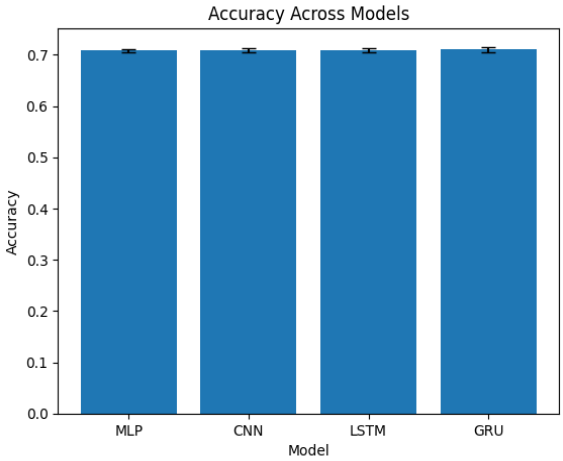

Figure 1. Accuracy Result

3.3 Precision Result

Precision evaluates the proportion of true positives among all predicted positives $\text{Precision} = \frac{TP}{TP+FP}$. GRU achieved the highest precision (0.5908 ± 0.0302), followed by CNN (0.5856 ± 0.0257) and LSTM (0.5824 ± 0.0215). The MLP model had the lowest precision (0.5803 ± 0.0175). Higher precision indicates fewer false positives, making GRU particularly suitable for tasks where over-predicting the positive class can lead to

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1310−1318
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6783

significant consequences, such as unnecessary treatments in healthcare settings. Precision is presented in the Figure 2.

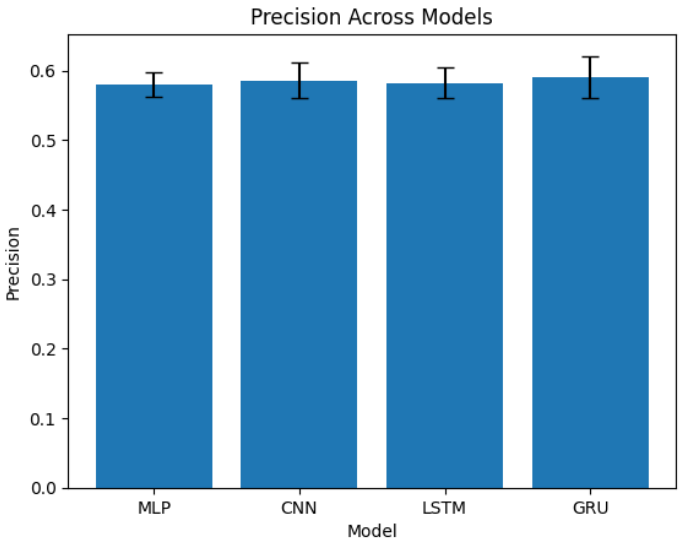

Figure 2. Precision Result

3.4 Recall Result

Recall measures the proportion of true positives identified among all actual positives: Recall $= \frac{\text{TP}}{\text{TP+FN}}$. CNN and LSTM recorded the highest recall values (0.2275 ± 0.0437 and 0.2274 ± 0.0330, respectively), outperforming GRU (0.2263 ± 0.0380) and MLP (0.2169 ± 0.0246). The consistently low recall values across all models highlight their difficulty in identifying minority-class samples, likely due to the class imbalance in the dataset. This limitation suggests the need for additional techniques, such as enhanced class balancing strategies or cost-sensitive learning, to improve the models' sensitivity. Recall is presented in the Figure 3.

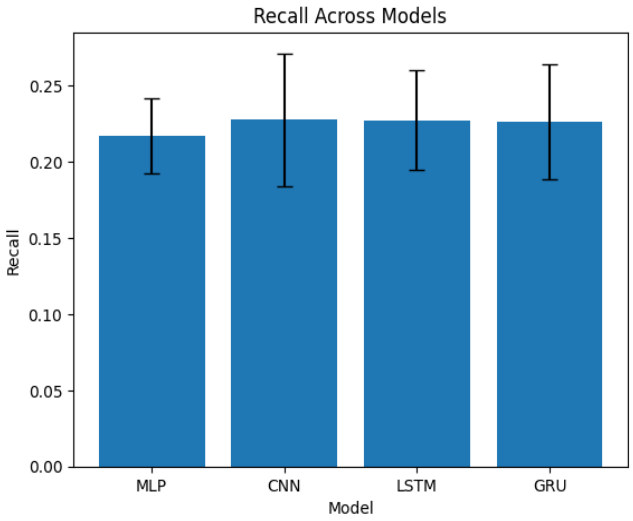

Figure 3. Recall Result

3.5 F1-Score and ROC-AUC Result

The F1-score is the harmonic mean of precision and recall F1 $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}}$ and the result is presented in the Figure 4. LSTM achieved the highest F1-score (0.3253 ± 0.0318), indicating a relatively better balance between precision and recall. CNN (0.3244 ± 0.0413) and GRU (0.3245 ± 0.0350) followed closely, while MLP recorded the lowest F1-score (0.3147 ± 0.0248). These findings suggest that recurrent architectures and CNNs are more effective in mitigating the trade-offs between precision and recall compared to traditional feedforward models like MLP. In addition, The ROC-AUC is presented in Figure 5. The metric evaluates the model's ability to distinguish between classes across various thresholds: ROC-AUC $= \int_0^1 TPR(FPR)\ d(FPR)$, where TPR and FPR denote the true positive rate and false positive rate, respectively. GRU recorded the highest ROC-AUC (0.5772 ± 0.0102), followed by CNN and LSTM, while MLP had the lowest ROC-AUC (0.5730 ± 0.0076). The higher ROC-AUC

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1310−1318
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6783

values of recurrent models and CNN demonstrate their ability to identify decision thresholds that maximize classification performance.

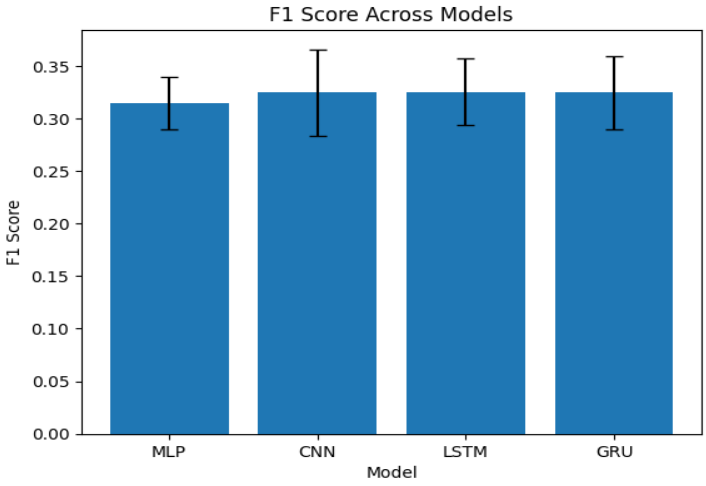

Figure 4. F1 Result

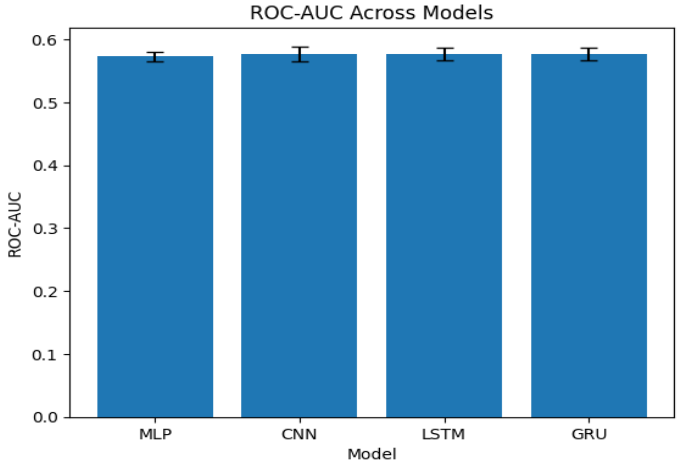

Figure 5. ROC-AUC Result

3.6 Discussion

The results emphasize the strengths and weaknesses of each DL model for HIV/AIDS classification. Recurrent architectures (LSTM and GRU) performed consistently well across most metrics, highlighting their ability to model temporal dependencies in the dataset. GRU's higher precision and lower variability suggest that its simpler gating mechanism is advantageous in handling the inherent complexity of the data. CNN demonstrated competitive performance, particularly in precision and F1-score, indicating its effectiveness in capturing spatial patterns within the dataset. However, the relatively lower recall across all models underscores the challenges posed by class imbalance. While SMOTE was applied during preprocessing, more sophisticated balancing techniques, such as generative adversarial networks for synthetic sample generation, could further enhance recall. MLP consistently underperformed, indicating its limited ability to capture hierarchical or temporal relationships in the data. This suggests that feedforward architecture may not be well-suited for tasks involving complex feature interactions and sequential dependencies. The overall results reveal that while GRU is the most promising model, all architecture shows limitations in recall. This shortcoming highlights the need for further enhancements, such as incorporating attention mechanisms or hybrid models like CNN-LSTM, which can leverage the strengths of multiple architectures.

3.7 Implications for Clinical Applications

The study's findings have significant implications for clinical practice. The higher precision of GRU and CNN models makes them suitable for screening tasks, where minimizing false positives is crucial. However, the low recall indicates that current models may miss a significant proportion of actual positive cases, potentially leading to underdiagnosis. Future research should focus on addressing these limitations through advanced model architecture, improved data preprocessing, and the integration of domain-specific knowledge. Specifically, incorporating attention-based mechanisms, such as transformers or hybrid CNN-Transformer architectures, could

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1310−1318
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6783

enhance the models' ability to capture long-range dependencies and complex feature interactions. These architectures have demonstrated superior performance in various medical classification tasks by selectively focusing on the most relevant input features. Additionally, improved data preprocessing techniques, such as feature engineering tailored to clinical markers or generative adversarial networks (GANs) for synthetic data augmentation, could help mitigate class imbalance and enhance model generalizability. Furthermore, integrating domain-specific knowledge, such as expert-driven feature selection or incorporating clinical decision rules into the modeling process, may enhance interpretability and diagnostic reliability. This could involve embedding medically relevant biomarkers into the training pipeline or utilizing hybrid models that combine deep learning with expert-defined heuristic rules.

3.8 Rationale for GRU's Superior Performance

The Gated Recurrent Unit (GRU) consistently outperformed the other models in this study due to its architectural efficiency and ability to handle sequential data effectively. GRU's simpler gating mechanism, compared to the Long Short-Term Memory (LSTM) model, reduces the number of parameters required while maintaining the capability to capture long-term dependencies. This reduction in complexity makes GRU less prone to overfitting and computationally more efficient, particularly for datasets with limited samples or high-dimensional features, as seen in this study. The superior precision achieved by GRU reflects its ability to minimize false positives effectively, which is crucial in healthcare diagnostics to prevent unnecessary treatments. Its lower variability across cross-validation folds indicates greater robustness, suggesting that GRU adapts well to variations in the training data. In contrast, LSTM, while capable of modeling long-term dependencies, may suffer from overfitting or increased sensitivity due to its more complex gating structure. GRU's ability to balance computational efficiency with robust performance makes it particularly suitable for clinical datasets, where efficiency, reliability, and scalability are critical. These strengths highlight why GRU emerged as the most effective model in this study. Future research could explore hybrid approaches, such as combining GRU with attention mechanisms or other architectures, to further enhance its predictive capabilities.

3.9 Advanced Analysis of Performance Metrics

The analysis of performance metrics highlights key trends and challenges in leveraging deep learning (DL) models for HIV/AIDS classification. Each model's behavior reflects unique strengths and limitations that are worth exploring further to provide actionable insights into their applicability. The results show that the GRU model consistently outperformed others in terms of accuracy (71.04%) and precision (59.08%), underscoring its effectiveness in maintaining reliable predictions and minimizing false positives. However, its recall (22.63%) was marginally lower than CNN and LSTM, indicating a need for enhanced sensitivity to detect minority-class samples.

LSTM achieved the highest F1-score (32.53%), which reflects a balanced trade-off between precision and recall. This suggests its ability to handle class imbalance better than MLP and CNN. Nevertheless, the gap between precision and recall indicates a potential trade-off between avoiding false positives and capturing all positive cases, which requires further investigation. Despite using SMOTE to address class imbalance, recall remained low across all models, with CNN and LSTM performing slightly better. This suggests that the synthetic samples generated by SMOTE may not fully represent the complexity of real-world minority-class samples. Future studies could employ generative adversarial networks (GANs) for data augmentation, which can create more realistic and diverse samples, potentially improving recall.

The recurrent models, specifically GRU and LSTM, demonstrated their ability to process sequential dependencies, as evidenced by their consistent performance across metrics. GRU stands out due to its simpler gating mechanism, which not only ensures computational efficiency but also maintains robust performance. This combination makes GRU an ideal candidate for large-scale clinical datasets where efficiency and reliability are paramount. Meanwhile, CNN showcased competitive performance, particularly in terms of its F1-score, which highlights its capability to capture spatial relationships within the data. This suggests that certain feature interactions in the dataset exhibit localized patterns that CNN is adept at exploiting effectively. On the other hand, the MLP model consistently underperformed across all metrics. Its inability to capture the temporal and hierarchical relationships critical for this classification task underscores the importance of selecting architectures that align with the structural properties of the data. Furthermore, an analysis of the standard deviations across metrics provides additional insights into model robustness. GRU displayed lower variability, particularly in accuracy and precision, which suggests its stable performance across different cross-validation folds. In contrast, CNN and LSTM exhibited higher variability in metrics such as recall and F1-score, reflecting their sensitivity to variations in the training data distribution. This observed variability underscores the need for further validation of these models on external datasets to ensure their generalizability and reliability in real-world applications.

3.10 Limitations of the Research

A significant challenge lies in the class imbalance within the dataset, as the minority class (individuals diagnosed with HIV/AIDS) is underrepresented. Despite employing the Synthetic Minority Oversampling Technique

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1310−1318
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6783

(SMOTE) to address this issue, the models still exhibited low recall values, indicating difficulty in effectively capturing patterns for minority-class samples. The generated synthetic samples may lack the nuanced variability of real-world data, which could have impacted the models' sensitivity. This suggests that a more balanced dataset, either through improved data collection or alternative augmentation strategies, could potentially enhance the models' ability to generalize to minority-class cases. Additionally, class imbalance may have influenced the optimization process, leading the models to prioritize overall accuracy rather than minority-class recall, further reducing sensitivity to critical cases. Another limitation is the scope of model architecture explored. Although four deep learning models (MLP, CNN, LSTM, and GRU) were evaluated, more advanced architectures, such as attention-based mechanisms or transformer models, were not included. These advanced models could potentially capture long-range dependencies and complex feature interactions more effectively, leading to improved performance. Future studies could explore these architectures to assess whether they provide better discrimination for the minority class. Furthermore, the models were validated using cross-validation on the same dataset, without testing on an external or independent dataset. While this approach is effective for initial evaluation, it raises concerns about the generalizability and robustness of the models in real-world clinical settings. The absence of external validation means that the models' true predictive power in diverse populations remains uncertain.

4. CONCLUSION

This study evaluated the performance of four deep learning models: Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU), in classifying HIV/AIDS status using clinical and laboratory data. The analysis was conducted across multiple evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to comprehensively assess the strengths and limitations of each model. The results revealed that GRU outperformed the other models in terms of accuracy (0.7104 ± 0.0052), precision (0.5908 ± 0.0302), and ROC-AUC (0.5772 ± 0.0102), demonstrating its robustness and effectiveness in handling sequential data. CNN and LSTM exhibited competitive performance, with LSTM achieving the highest F1-score (0.3253 ± 0.0318), highlighting its ability to balance precision and recall. However, MLP consistently underperformed, indicating that feedforward architectures are less suited for complex datasets with temporal and hierarchical feature dependencies. Despite these findings, a common challenge across all models was the low recall, reflecting difficulties in identifying minority-class samples. This limitation underscores the need for further research to address class imbalance through advanced data balancing techniques, such as generative adversarial networks (GANs) or cost-sensitive learning. The integration of hybrid architectures, such as CNN-LSTM or attention-based models, also holds promise for enhancing feature extraction and improving overall classification performance. The implications of this study extend beyond model evaluation, emphasizing the importance of precision in healthcare diagnostics to minimize false positives while highlighting the critical need to improve sensitivity to avoid underdiagnosis. GRU, with its stability and superior overall performance, emerges as a strong candidate for real-world deployment in HIV/AIDS screening systems. However, the findings also stress the necessity for additional validation on external datasets to ensure generalizability and robustness in diverse clinical settings. Future directions for this research include the exploration of transformer-based architecture and ensemble methods to further enhance performance. Additionally, incorporating explainability techniques to interpret model predictions will be essential for building trust and facilitating adoption in clinical environments. By addressing these challenges, deep learning models can achieve greater accuracy, fairness, and reliability, ultimately contributing to improved patient outcomes and advancing the application of artificial intelligence in healthcare.

ACKNOWLEDGMENT

The authors would like to express their deepest gratitude to Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia, for the invaluable support and resources provided throughout this research. The institution's commitment to fostering academic excellence and innovation has been instrumental in the successful completion of this study. We are particularly thankful for the access to computational facilities, research materials, and the supportive environment that enabled us to carry out this work effectively. This research stands as a testament to the university's dedication to advancing knowledge and contributing to meaningful scientific endeavors.

REFERENCES

[1] R. B. Sonawane and G. D. Barkade, “Literature Review on Acquired Immunodeficiency Syndrome (AIDS).,” Syst. Rev. Pharm., vol. 14, no. 5, 2023.
[2] E. Kokori et al., “Implications of long-acting antiretrovirals (LAARVs) for HIV treatment in Sub-Saharan Africa,” Discov. Public Heal., vol. 21, no. 1, pp. 1–7, 2024.
[3] J. Yang, X. Zheng, S. Zhang, and H. Wang, “Epidemiology-based Analysis of Characteristics of Dual Infection of Tuberculosis/Acquired Immune Deficiency Syndrome (AIDS) and Drug Resistance Mechanism of Related Genes,” Cell. Mol. Biol., vol. 68, no. 2, pp. 109–118, 2022.

Journal of Information System Research (JOSH)
Volume 6, No. 2, Januari 2025, pp 1310−1318
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v6i2.6783

[4] Jocelyn et al., "HIV/AIDS in Indonesia: current treatment landscape, future therapeutic horizons, and herbal approaches," Front. Public Heal., vol. 12, p. 1298297, 2024.
[5] E. Kumah, D. S. Boakye, R. Boateng, and E. Agyei, "Advancing the global fight against HIV/Aids: Strategies, barriers, and the road to eradication," Ann. Glob. Heal., vol. 89, no. 1, 2023.
[6] F. E. T. Foka and H. T. Mufhandu, "Current ARTs, virologic failure, and implications for AIDS management: a systematic review," Viruses, vol. 15, no. 8, p. 1732, 2023.
[7] S. Qiao, X. Li, B. Olatosi, and S. D. Young, "Utilizing Big Data analytics and electronic health record data in HIV prevention, treatment, and care research: a literature review," AIDS Care, vol. 36, no. 5, pp. 583–603, 2024.
[8] Y. Li et al., "The predictive accuracy of machine learning for the risk of death in HIV patients: a systematic review and meta-analysis," BMC Infect. Dis., vol. 24, no. 1, p. 474, 2024.
[9] R. Saha, L. Malviya, A. Jadhav, and R. Dangi, "Early stage HIV diagnosis using optimized ensemble learning technique," Biomed. Signal Process. Control, vol. 89, p. 105787, 2024.
[10] A. Rehman, S. Naz, and I. Razzak, "Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities," Multimed. Syst., vol. 28, no. 4, pp. 1339–1371, 2022.
[11] 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.
[12] A. Zhang, L. Xing, J. Zou, and J. C. Wu, "Shifting machine learning for healthcare from development to deployment and from models to data," Nat. Biomed. Eng., vol. 6, no. 12, pp. 1330–1345, 2022.
[13] A. Velu, "AIDS Virus Infection Prediction." Kaggle, 2023.
[14] S. N. Mohsin et al., "The role of artificial intelligence in prediction, risk stratification, and personalized treatment planning for congenital heart diseases," Cureus, vol. 15, no. 8, 2023.
[15] M. Athar, "Potentials of artificial intelligence in familial hypercholesterolemia: Advances in screening, diagnosis, and risk stratification for early intervention and treatment," Int. J. Cardiol., vol. 412, p. 132315, 2024.
[16] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," Information, vol. 15, no. 4, p. 235, 2024.
[17] E. S. Mohamed, T. A. Naqishbandi, S. A. C. Bukhari, I. Rauf, V. Sawrikar, and A. Hussain, "A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms," Healthc. Anal., vol. 3, p. 100185, 2023.
[18] A. Makandar and M. N. Jadhav, "Disease Recognition in Medical Images Using CNN-LSTM-GRU Ensemble, a Hybrid Deep Learning," in 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2023, pp. 1–9.
[19] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," Information, vol. 15, no. 9, p. 517, 2024.
[20] Z. Tarek et al., "An optimized model based on deep learning and gated recurrent unit for COVID-19 death prediction," Biomimetics, vol. 8, no. 7, p. 552, 2023.
[21] M. A. Basri, "Evaluating the Usefulness of Synthetic Data in Healthcare: Applications in Predictive Modeling and Privacy Protection," University of Waterloo, 2024.
[22] S. N. Manivannan, C. D. Arenas, N. D. Grubaugh, and C. B. Ogbunugafor, "The importance of epistasis in the evolution of viral pathogens," Evolution (N. Y)., vol. 20, p. 23.
[23] E. G. Anderson, D. R. Keith, and J. Lopez, "Opportunities for system dynamics research in operations management for public policy," Prod. Oper. Manag., vol. 32, no. 6, pp. 1895–1920, 2023.