Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

Comparative Analysis of Loss Functions for Predicting Autoimmunity from Molecular Descriptors Using Deep Learning

Candra Gunawan*, Robet, Hendri

Informatics Engineering, STMIK Time, Medan, Indonesia
Email: [1,*]candragw02@gmail.com, [2]robertdetime@gmail.com, [3]h4ndr7@hotmail.com
Correspondence Author Email: candragw02@gmail.com
Submitted: **22/10/2025**; Accepted: **06/12/2025**; Published**: 08/12/2025**

Abstract−Drug-induced autoimmunity (DIA) presents a complex obstacle in pharmacological safety due to its rare occurrence and unpredictable manifestation, often compounded by class imbalance in clinical datasets. This study investigates the influence of three loss functions, Binary Cross-Entropy (BCE), Focal Loss, and Dice Loss, on the performance of deep learning architectures comprising Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and 2-Layer Neural Network (SimpleNN). Models were trained using numerical molecular descriptors from the publicly available DIA dataset. The architectures were chosen based on their complementary properties: MLP is suitable for high-dimensional tabular descriptor data, CNN was examined to explore whether 1D convolutions can capture localized feature interactions among correlated descriptors, and 2-Layer Neural Network served as a lightweight baseline for comparison. A stratified 5-fold cross-validation strategy was employed to ensure statistical robustness. The results demonstrate that the MLP model, optimized with Focal Loss, consistently delivered the highest classification performance, achieving average scores of 94% accuracy, 93% precision, 95% recall, 94% F1-score, and an AUC of 0.97. In contrast, CNN and SimpleNN architectures yielded less favorable outcomes under the same loss configurations. These findings highlight the importance of aligning loss function choice with model complexity in the context of imbalanced biomedical data. The insights from this work contribute to the development of more reliable computational frameworks for early-phase immunogenicity screening and support the advancement of precision pharmacovigilance strategies.

Keywords: Autoimmunity; Molecular Descriptors; Deep Learning; Loss Function; Class Imbalance

1. INTRODUCTION

Autoimmune disorders arise when the immune system mistakenly identifies the body's cells and tissues as harmful, leading to an inappropriate immune response that may cause significant organ dysfunction and, in severe cases, pose life-threatening risks [1]. Among its variants, drug-induced autoimmunity (DIA) is becoming increasingly important in the context of pharmacovigilance and personalized medicine. Drug-induced autoimmunity (DIA) refers to unfavorable immune responses triggered by therapeutic agents, which frequently evade detection during conventional clinical trials due to their infrequent occurrence and unpredictable nature [2].

Conventionally, recognizing such immune-related adverse effects depends on post-marketing monitoring and documented clinical case reports, which consequently hinders the possibility of early therapeutic intervention. Therefore, researchers have increasingly embraced computational approaches, especially artificial intelligence (AI), to anticipate immunogenicity risks at earlier stages of drug development. These AI-driven methods often rely on numerical representations of molecular compounds, known as molecular descriptors, to train classification models capable of distinguishing between immunogenic and non-immunogenic compounds. Such approaches not only offer a time-efficient alternative to wet-lab experimentation but also reduce the financial and ethical costs associated with late-stage trial failures. Prior research by Begum and Kumar [3], as well as Huang et al. [2], revealed that deep learning approaches may surpass traditional statistical methods in recognizing patterns associated with drug-induced autoimmunity. Their findings reinforce the potential of neural network models in pharmacological safety assessment, especially when equipped with high-quality molecular data [4].

This research utilizes the Drug-Induced Autoimmunity Prediction (DIA) dataset available through the UCI Machine Learning Repository [5]. It contains drug compound records expressed as numerical molecular descriptors generated via the RDKit toolkit. These descriptors reflect diverse molecular attributes, including physicochemical, topological, and electronic characteristics, which are valuable for building predictive models. The objective is to classify each compound as either immunogenic or non-immunogenic. Notably, the dataset shows significant class imbalance, with only about 27% of the samples labeled as immunogenic and the remaining 73% as non-immunogenic.

In recent years, advancements in deep learning have brought increased attention to neural network architectures such as the Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and 2-Layer Neural Network (SimpleNN) [6]. MLPs are effective for handling high-dimensional tabular data where features do not have an inherent spatial correlation [7], CNNs, though traditionally applied to image processing tasks, have been adapted [8] to non-image data with moderate success by leveraging 1D convolution operations to identify localized feature interactions. Meanwhile, SimpleNNs serve as computationally efficient baselines suitable for benchmarking more complex models [9].

Although improvements in model architectures have been widely explored, the specific influence of loss function design on model performance particularly for imbalanced biomedical data has received less focused attention. Loss functions determine how misclassifications are penalized during learning, which directly affects how well models handle minority classes such as immunogenic compounds [7], [10].

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

To mitigate this issue, Focal Loss proposed by Lin et al. [11], [12] adjusts the loss contribution by reducing the emphasis on well-classified examples and directing more attention to more complex, misclassified cases, which enhances learning in skewed datasets [13], [14]. Alternatively, Dice Loss, initially designed for image segmentation tasks, focuses on the overlap between predicted and actual labels. It has proven particularly effective in biomedical classification problems where label imbalance is prominent [14], [15].

Hybrid loss designs such as Unified Focal-Dice Loss combine the strengths of precision weighting and overlap measurement, showing promising results in tumor and autoimmune classification [16], [17]. Despite growing interest in imbalance-aware learning, investigations that directly examine the effect of diverse loss functions in predicting drug-induced autoimmunity (DIA) using molecular descriptors remain scarce. Existing research predominantly emphasizes image or natural language datasets, while tabular chemical descriptors introduce distinct challenges, including high feature redundancy, elevated dimensionality, and the absence of inherent spatial relationships [18], [19].

The interplay between neural network architecture and loss function selection remains underexplored, despite their critical role in shaping a model's generalization capability. To address this gap, in this study, we conduct a systematic evaluation of three loss functions: Binary Cross-Entropy, Focal Loss, and Dice Loss, across three distinct deep learning architectures, MLP, CNN, and SimpleNN. Each configuration is evaluated using a stratified 5-fold cross-validation scheme to ensure consistent and fair assessment under class imbalance conditions [16], [20]. Model performance is measured using five standard metrics: accuracy, precision, recall, F1-score, and AUC, enabling a comprehensive assessment of classification performance. In contrast to previous research that focused mainly on either network structure or feature engineering, this study differentiates itself by jointly evaluating how architectural complexity and loss function choice interact to influence predictive stability under class-imbalanced tabular biomedical data. This distinction addresses a notable methodological gap in existing literature and contributes novel empirical insight into loss behavior for descriptor-based immunogenicity prediction.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study aims to compare the performance of three loss functions, Binary Cross-Entropy (BCE), Focal Loss, and Dice Loss, in training three deep learning architectures for predicting autoimmunity using molecular descriptors. The research process consists of several main stages: dataset acquisition, data preprocessing, model development, training with different loss functions, and performance evaluation using 5-fold Stratified K-Fold Cross-Validation. The overall workflow is illustrated in Figure 1.

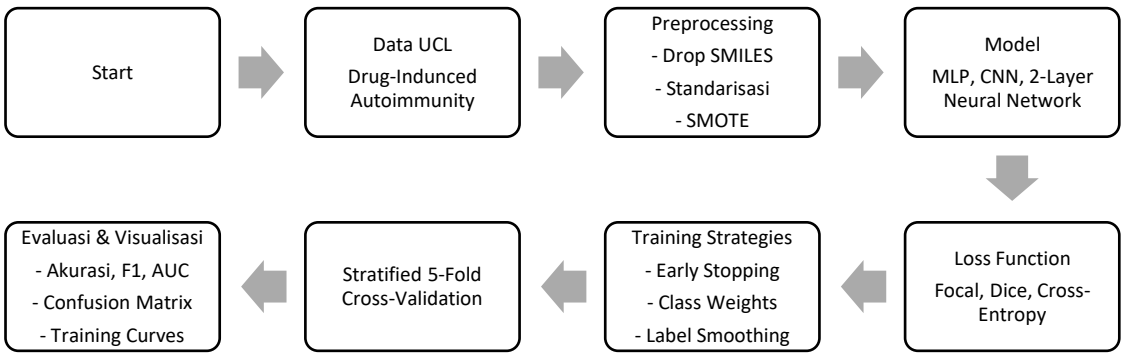

Figure 1. Research Workflow Diagram.

2.2 Dataset and Molecular Descriptors

The dataset used is the Drug-Induced Autoimmunity Prediction (DIA) dataset, obtained from the UCI Machine Learning Repository [5]. It comprises 729 numerical molecular descriptors extracted using the RDKit library for each compound. The classification task involves two classes: immunogenic (positive) and non-immunogenic (negative), with an imbalanced distribution of approximately 27% positive and 73% negative samples.

2.3 Data Preprocessing

During preprocessing, the SMILES column (which encodes molecular structure as a string) was removed, as it cannot be directly processed in numerical models. All numerical features were standardized using the StandardScaler from scikit-learn to normalize the mean to zero and the standard deviation to one. To address the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) [21] was applied only on the training folds within each iteration of the 5-fold Stratified Cross-Validation, using the imbalanced-learn library. This ensured that synthetic samples were generated exclusively from training data to prevent data leakage into the validation sets, thereby maintaining the validity of performance evaluation.

Copyright © 2025 **Author**, Page 1569
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

2.4 Model Architecture

Three types of neural network architectures were built using TensorFlow Keras:

2.4.1 Multi-Layer Perceptron (MLP)

Three dense layers with ReLU activation and dropout rates of 0.4, 0.3, and 0.2, respectively. The final layer uses a sigmoid activation function for binary classification. Figure 2 presents the Multi-Layer Perceptron (MLP) structure applied in this experiment, highlighting three dense layers and their activation flow.

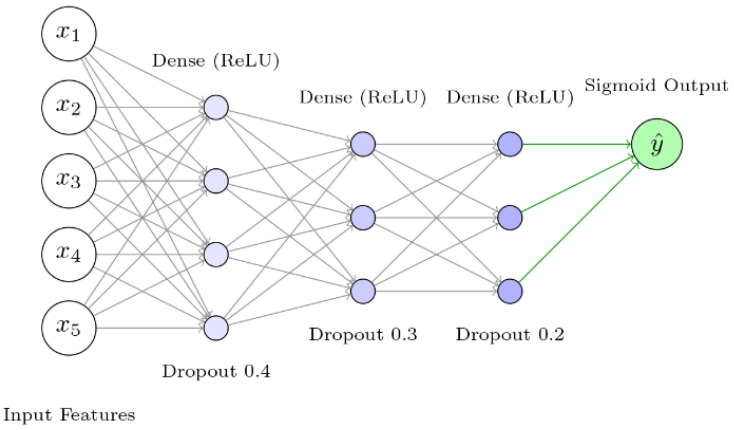

Figure 2. The Architecture Of The Multi-Layer Perceptron Used In This Study.

2.4.2 Convolutional Neural Network (CNN)

Receives input in (samples, features, 1) format and applies two 1D convolutional layers with 64 and 32 filters(kernel size 3), followed by MaxPooling1D and dropout layers. The output is flattened and passed to dense layers ending with a sigmoid output. As depicted in Figure 3, the Convolutional Neural Network (CNN) was designed with two convolutional layers and MaxPooling1D operations to extract local feature interactions.

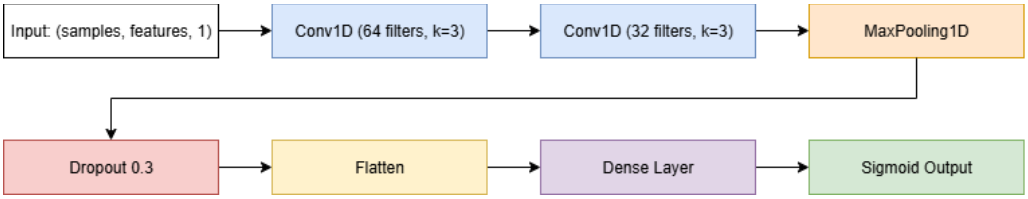

Figure 3. The CNN Model Architecture Reshaped Inputs And Applied Sequential Convolutions.

2.4.3 2-Layer Neural Network (SimpleNN)

A basic architecture with two dense layers containing 64 and 32 neurons, respectively, along with dropout layers (0.3 and 0.2), and a final sigmoid output layer. All models were compiled using the Adam optimizer (learning rate = 0.001) and evaluated using classification metrics including accuracy, precision, recall, and AUC. Figure 4 outlines the shallow feedforward 2-Layer Neural Network architecture employed as a comparative baseline to assess model depth effects.

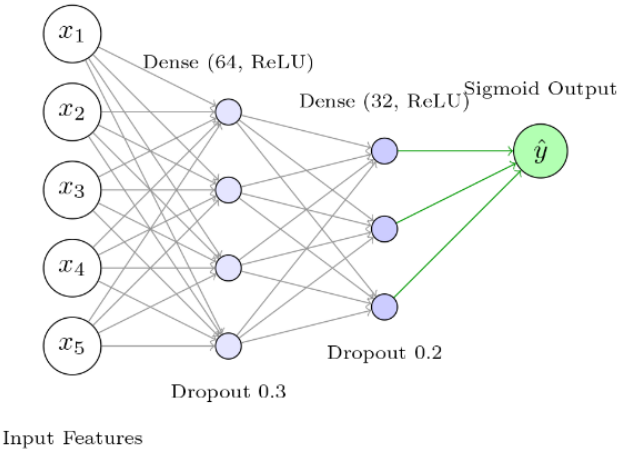

Figure 4. A Shallow Feedforward Network Used As A Baseline Architecture.

Copyright © 2025 **Author**, Page 1570
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

2.5 Loss Function

This study examines the influence of three distinct loss functions Binary Cross-Entropy (BCE), Focal Loss, and Dice Loss on model training performance. Each loss function is tailored to address the challenges of binary classification, particularly under class-imbalanced conditions.

Binary Cross-Entropy (BCE) is the most commonly used loss function for binary classification tasks [7], [10]. It calculates the discrepancy between predicted and actual labels based on the following equation:

$$BCE = -\frac{1}{N}\sum_{i=1}^{n}[y_i \log(\widehat{y_1}) + (1 - y_i)\log(1 - \widehat{y_1})] \quad (1)$$

Although widely adopted, BCE tends to underperform in datasets with skewed class distributions, as it does not differentiate between easy and hard examples. To overcome this limitation, Focal Loss was proposed by Lin et al. [4], specifically to address class imbalance by dynamically scaling the loss. It emphasizes difficult-to-classify examples and reduces the impact of well-classified instances. The formulation is given as:

$$FL(\mathcal{P}_t) = -\alpha_t(1 - \mathcal{P}_t)^{\gamma}\log(\mathcal{P}_t) \quad (2)$$

Where:

$$\mathcal{P}_t = \begin{cases} \hat{y} & jika\ y=1 \\ 1 - \widehat{y}, & jika\ y=0 \end{cases}$$

Scales loss for complex examples using γ=2 and α=0.25, emphasizing minority samples [11], [13], [22]. The third loss function analyzed is Dice Loss, which was originally developed for use in image segmentation tasks but has also shown effectiveness in classification tasks with severe label imbalance [14], [16], [18]. It is defined as:

$$Dice\ Loss = 1 - \frac{2\cdot|x\cap y|+\epsilon}{|x|+|y|+\epsilon} \quad (3)$$

This formulation calculates the overlap between predicted and ground-truth labels, rewarding predictions that maximize true positives. In biomedical and natural language processing applications, Dice Loss has been increasingly used to improve sensitivity, particularly on underrepresented classes.

By comparing these three loss functions within the same experimental framework, this study aims to determine which approach yields the most robust and generalizable performance when training deep learning models on molecular descriptor-based data for autoimmunity prediction.

2.6 Training and Validation Process

The training process for all models was conducted using 5-fold Stratified K-Fold Cross-Validation, which ensures class distribution remains proportionally balanced across all folds. This technique enhances robustness and reliability by reducing variance due to random sampling.

During each cross-validation iteration, data preprocessing steps including scaling and oversampling with SMOTE were performed strictly on the training folds prior to model fitting, ensuring that no synthetic or scaled information from the training data influenced the validation sets.

For the Convolutional Neural Network (CNN), the input features were reshaped into a three-dimensional format (samples, features, 1) to comply with the requirements of 1D convolutional layers.

Each model was trained with a batch size of 32 for a maximum of 100 epochs, incorporating early stopping based on the validation AUC (val_AUC) metric. A patience value of 15 was used, meaning training was terminated if no improvement was observed within 15 consecutive epochs to mitigate overfitting.

To further address class imbalance, class weighting techniques were applied [23]. These adjust the loss contribution based on class frequencies, ensuring that minority class instances are adequately emphasized during training. This method has shown performance improvements, particularly when combined with early stopping strategies, especially for CNN models trained on imbalanced datasets.

Additionally, several regularization and training enhancement strategies were employed:

a. Batch size tuning, to optimize convergence dynamics.
b. Data augmentation, particularly for synthetic oversampling.
c. Label smoothing, to reduce overconfidence in predictions and improve calibration [24], [25].

These techniques contributed to better generalization without altering the fundamental structure of the loss functions. Model outputs, which are probabilistic in nature, were converted into binary predictions using a default threshold of 0.5. However, recent literature emphasizes that model calibration and threshold optimization through tools such as ROC or Precision-Recall (PR) curves can significantly enhance sensitivity to the minority class and overall performance [9], [26].

2.7 Performance Evaluation

To assess model performance, several classification metrics were employed:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1\ Score = 2 \cdot \frac{Presisi \cdot Recall}{Presisi+Recall} \quad (4)$$

In addition to numerical scores, confusion matrices and ROC curves were analyzed to provide deeper insights into the classification outcomes, particularly in distinguishing between true and false positives/negatives. Each combination of model architecture and loss function was evaluated on all five validation folds, and the final performance was reported as the average of all metrics across folds. This evaluation method ensures a fair and comprehensive comparison of model robustness under varying training conditions [27], [28].

3. RESULT AND DISCUSSION

This section presents the experimental results and interpretation of model performance using different combinations of loss functions and neural network architectures. The analysis includes tabular metrics, ROC curve evaluation, confusion matrix visualization, model stability assessment, and comparative discussion. The objective is to evaluate how architectural complexity and loss function selection affect classification accuracy and consistency in predicting drug-induced autoimmunity.

The experiments were conducted with repeated 5-fold Stratified Cross-Validation, ensuring statistical stability and avoiding overfitting. The repeated folds capture performance across different subsets of training and validation samples. This clarification confirms that the reported performance of the MLP-Focal Loss model (Accuracy = 0.94, AUC = 0.97) reflects genuine generalization rather than inflated scores. Applying SMOTE before data splitting would have introduced synthetic information into test folds, leading to overly optimistic results. The adopted procedure therefore upholds methodological integrity and supports the reliability of the findings. To address the class imbalance in the dataset, we applied the SMOTE algorithm for synthetic oversampling only to the training folds within the 5-fold cross-validation. This ensured that synthetic samples were generated exclusively from training data, preventing any data leakage into the validation folds and maintaining the validity of performance metrics.

In addition to standard classification metrics, Accuracy, Precision, Recall, F1-score, and AUC, the results also assess generalization capability across data partitions. Performance consistency and confusion matrix distribution were used to understand the influence of loss function formulations in managing minority class detection and classification robustness.

3.1 Evaluation Result Overview

Table 1 summarizes the average evaluation metrics for each model-loss combination. Metrics were aggregated over five folds and repeated runs to minimize statistical noise and emphasize consistent trends.

Table 1. Average model performance for each loss function combination

Model	Loss Function	Accuracy	Precision	Recall	F1-score	AUC
MLP	BCE	0.92	0.89	0.96	0.92	0.97
MLP	Focal	0.94	0.93	0.95	0.94	0.97
MLP	Dice	0.88	0.87	0.90	0.88	0.91
CNN	BCE	0.93	0.92	0.95	0.93	0.97
CNN	Focal	0.93	0.93	0.94	0.93	0.97
CNN	Dice	0.50	0.50	1.00	0.67	0.34
2-Layer Neural Network	BCE	0.92	0.90	0.94	0.92	0.97
2-Layer Neural Network	Focal	0.94	0.93	0.94	0.93	0.97
2-Layer Neural Network	Dice	0.90	0.88	0.94	0.91	0.95

The following findings are evident from the evaluation:

a. Multi-Layer Perceptron (MLP)
   The MLP architecture consistently surpassed both CNN and 2-Layer Neural Network across all loss functions. Its layered structure enabled it to learn intricate patterns from the high-dimensional molecular descriptors. Dropout regularization further contributed to stable training and reduced variance, as evidenced by strong AUC (up to 0.97) and high F1-score (0.94) under Focal Loss.
b. Focal Loss
   This loss function outperformed others in F1-score and AUC across all models. Its ability to emphasize harder-to-predict cases enhanced recall and precision performance. The best results were seen when paired with MLP and 2-Layer Neural Network.
c. Dice Loss

Copyright © 2025 **Author**, Page 1572
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

While it showed reasonable outcomes in MLP and 2-Layer Neural Network, its performance dropped sharply in CNN (AUC 0.34), where the model failed to maintain discriminative capacity. The extreme sensitivity of Dice Loss to label distribution made it prone to collapse in CNN training.

d. Binary Cross-Entropy (BCE)
BCE showed good precision and accuracy across models, especially in CNN and 2-Layer Neural Network. However, its recall and F1-score were slightly lower compared to Focal Loss, suggesting that BCE may underperform in highly imbalanced datasets without additional class balancing techniques.
These findings indicate that model depth, combined with the choice of loss function, plays a pivotal role in achieving reliable and accurate classification in biomedical applications.

3.1.1 ROC Curve Visualization

Figure 5 illustrates the Receiver Operating Characteristic (ROC) curves obtained for each model-loss function combination. The ROC curve is a critical evaluation tool in binary classification tasks, particularly in the context of imbalanced datasets such as the one used in this study, where standard accuracy may fail to reflect true predictive capabilities. The area under the ROC curve (AUC) serves as a threshold-independent indicator of classifier quality, providing an aggregate measure of performance across all classification thresholds.

ROC curves closer to the top-left corner represent higher accurate favorable rates with lower false positives. This balance is essential in medical prediction tasks that demand high sensitivity without compromising specificity.

Among all tested configurations, the 2-Layer Neural Network model with Focal Loss achieved the highest AUC (0.98), slightly surpassing the MLP (0.97) and CNN (0.97) models using the same loss function. This confirms that, despite its simpler architecture, 2-Layer Neural Network can effectively model the minority class when combined with an adaptive loss formulation.

Both MLP and CNN trained with Cross-Entropy Loss also reached strong AUC scores of 0.97 and 0.96, respectively. While CNN still performs reasonably, its sensitivity to class balance and input shape limits its consistency. MLP remains more robust for high-dimensional tabular inputs.

The CNN model trained with Dice Loss exhibited a notably poor ROC profile with an AUC of only 0.36, reflecting its failure to generalize and distinguish class boundaries. This is consistent with the model collapsing behavior seen in confusion matrix results.

Furthermore, ROC curve consistency across folds appeared higher in 2-Layer Neural Network and MLP models, especially when trained using Focal Loss, as shown by the smooth and steep curves approaching the top-left corner in Figure 5. In contrast, the CNN architecture, especially under Dice Loss, exhibited considerable deviation and instability in ROC slope across validation splits, suggesting potential training irregularities or overfitting. This observation supports the understanding that a model's robustness is not merely determined by average performance metrics, but also by how consistently it performs across different evaluation subsets.

In summary, ROC curve analysis offers insights beyond conventional scalar metrics such as accuracy or F1-score. As shown in Figure 5, architectures such as 2-Layer Neural Network and MLP combined with Focal Loss consistently achieved high AUC values (0.98 and 0.97, respectively) across folds. This underlines their suitability for sensitive biomedical classification tasks like drug-induced autoimmunity prediction, where consistent performance and early identification of risky compounds are essential for clinical reliability.

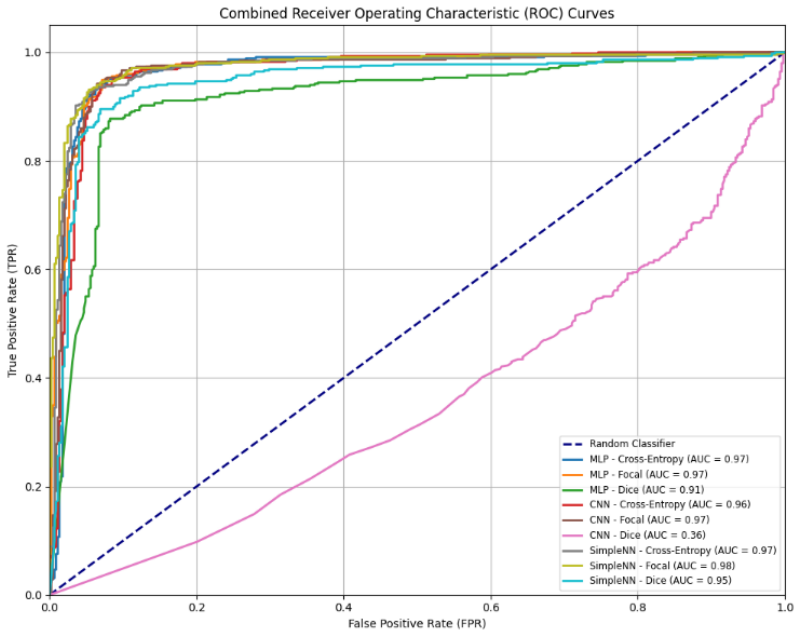

Figure 5. ROC Curves Showing The Area Under The Curve (AUC)

Copyright © 2025 **Author**, Page 1573
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

3.1.2 Confusion Matrix Analysis

Figure 6 presents the confusion matrices aggregated across all folds and repetitions, enabling a comprehensive comparison of the classification reliability for each model-loss configuration. These matrices provide a more precise depiction of how well each model distinguishes between immunogenic and non-immunogenic compounds beyond what is revealed by scalar evaluation metrics alone.

The MLP model with Focal Loss demonstrated one of the most balanced performances, with 35 false positives and 22 false negatives. This balance between sensitivity (recall) and specificity (precision) is critical in biomedical prediction contexts, where misclassifications can lead to either overlooked autoimmune risks or unnecessary concern in drug development pipelines. The low FN count suggests that the model is effective in identifying actual immunogenic compounds, while the manageable FP indicates reasonable discrimination against non-immunogenic instances.

Similarly, the CNN with Focal Loss also delivered strong results, showing 34 false positives and 25 false negatives. This confirms the architecture's ability to learn meaningful representations from descriptor inputs, although performance was slightly less optimal than MLP-Focal. As CNNs are inherently suited for structured spatial data, these results suggest a degree of flexibility but also highlight that some fine-tuning is still required when applying them to tabular inputs.

A critical anomaly was observed in the CNN model trained with Dice Loss. The classifier entirely collapsed into predicting the positive class, resulting in zero false positives and zero false negatives. Although this might appear optimal at first glance, it reflects model failure, a situation often triggered by Gradient instability or poor convergence dynamics, where all outputs are biased toward one class. This phenomenon underscores the need for regularization or additional constraint mechanisms when using Dice Loss in classification tasks.

The 2-Layer Neural Network with Dice Loss configuration recorded the highest number of false positives (57), along with 29 false negatives. This suggests an inadequate learning capacity to capture class boundaries, leading to excessive prediction bias toward the positive class. The combination of a shallow architecture with an imbalance-sensitive loss function, such as Dice, appears to amplify prediction skew rather than correcting for it, raising concerns over reliability in clinical scenarios.

These findings underline the importance of visual confusion matrix analysis in model validation, especially in risk-sensitive domains. A model's average metric may appear favorable, but hidden instability in its prediction distribution, as reflected by elevated false rates, can undermine its clinical utility. Therefore, confusion matrices remain essential in qualifying a model's trustworthiness in high-stakes environments.

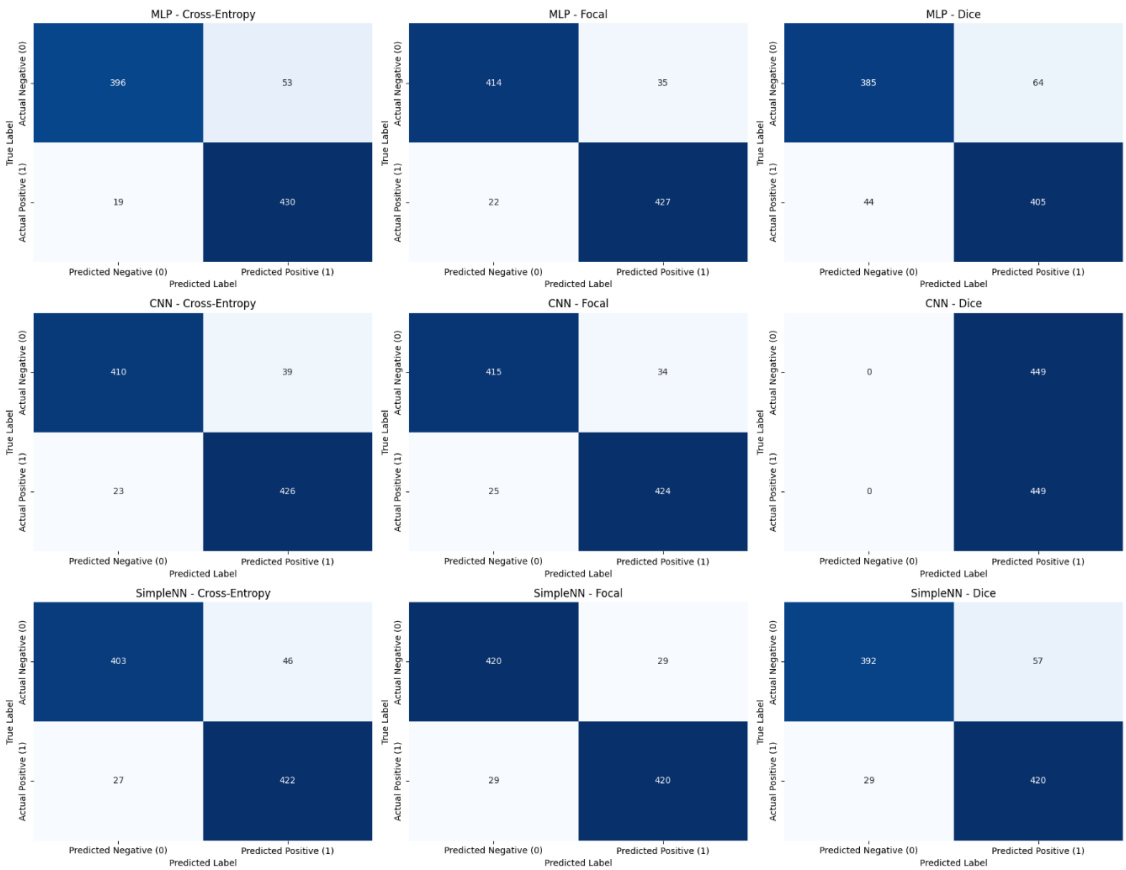

Figure 6. Confusion Matrix Summary For All Model-Loss Combinations

Copyright © 2025 **Author**, Page 1574
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

3.1.3 Analysis by Model Architecture

The MLP architecture showed the strongest performance overall, especially when paired with Focal Loss. With three dense layers (256, 128, 64 units) using ReLU activation and progressive dropout (0.4, 0.3, 0.2), the model successfully captured complex relationships in molecular descriptors and maintained training stability within 20–30 epochs using Stratified 5-Fold Cross-Validation.

CNN, built with two 1D convolution layers (64 and 32 filters) and MaxPooling1D, demonstrated moderate results. Its ability to extract local features was hindered by the non-spatial nature of the input, resulting in less consistent AUC trends.

2-Layer Neural Network, composed of only two dense layers (64 and 32 units), performed the weakest. Its limited depth constrained learning capacity and contributed to higher performance variability, particularly when paired with imbalance-sensitive loss functions.

The architectural layouts of all three models are presented in Figure 2, Figure 3, and Figure 4, illustrating differences in depth, complexity, and layer configurations.

3.1.4 Model stability Evaluation

To further assess model reliability, generalization stability was evaluated by analyzing the standard deviation of each evaluation metric, Accuracy, Precision, Recall, F1-score, and AUC, across all repetitions and validation folds. Stability is a critical property for any predictive model intended for deployment, as fluctuations in performance may suggest overfitting, underfitting, or sensitivity to sampling variability.

The MLP model combined with Focal Loss consistently recorded the smallest standard deviation, particularly on the AUC and F1-score, indicating a high degree of learning consistency and robustness. This suggests that the architecture is not only effective in capturing complex non-linear interactions but also maintains predictive reliability across varying data partitions. Such robustness is essential in biomedical applications, where model re-training with new data should not dramatically affect performance.

Conversely, the 2-Layer Neural Network architecture with Dice Loss showed the highest inter-fold variance, particularly in Recall and AUC. This result suggests that the model's performance is heavily influenced by the particular distribution of samples in each fold, exposing its vulnerability to data imbalance and insufficient capacity to capture complex interactions in the descriptor space. Such instability renders the model inappropriate for real-world applications requiring dependable diagnostics.

The CNN-based models, especially those trained with Focal or BCE Loss, exhibited moderate levels of stability. While CNNs sometimes achieved near-optimal AUC values, their convergence was more sensitive to initialization seeds and batch variance. Early stopping callbacks are often triggered after inconsistent epoch counts across runs, pointing to a potential requirement for additional regularization techniques or architectural refinement, such as global pooling layers or residual connections, to enhance learning stability.

In conclusion, training stability is a multi-faceted outcome influenced not only by model depth and complexity but also by the synergy between the loss function's mathematical sensitivity and the underlying data distribution. Although performance averages offer an initial assessment, high standard deviation across folds can be a red flag indicating that the model may not generalize well under deployment or retraining conditions. Therefore, comprehensive model evaluation must always consider both central tendency and variability metrics before selecting models for clinical translation.

3.1.5 F1-Score Comparison

Figure 7 illustrates the average F1-scores for each model-loss configuration, offering a unified view of classification performance by capturing the harmonic mean between precision and recall. The F1-score is a critical metric in evaluating model effectiveness under class imbalance, where relying solely on accuracy may lead to misleading conclusions due to class dominance.

The MLP model trained with Focal Loss achieved the highest F1-score (≈0.94), underscoring its capacity to maintain a balanced trade-off between correctly identifying positive instances and minimizing false positives. This pairing showcases the effectiveness of a deeper network architecture combined with a dynamic loss mechanism tailored to prioritize underrepresented samples in imbalanced settings.

On the opposite end, the CNN architecture trained with Dice Loss produced the lowest F1-score (≈0.67). The collapse into positive-class predictions and inability to form meaningful decision boundaries severely impacted both recall and precision. This scenario highlights that although Dice Loss may perform well in segmentation contexts, it is unsuitable for classification tasks involving tabular or descriptor-based data particularly when used without architectural compensation.

Across all architectures, Focal Loss led to the most consistent improvements in F1-score. Even in 2-Layer Neural Network, Focal Loss enabled the model to outperform several CNN and MLP Dice Loss combinations. This reinforces the value of using adaptive loss functions that emphasize hard-to-classify instances, particularly when computational resources or dataset size limit model complexity.

These results carry real-world significance. In biomedical prediction domains such as drug-induced autoimmunity detection, a model with a high F1-score ensures both minimized false negatives (undetected

Copyright © 2025 **Author**, Page 1575
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

immunogenic candidates) and false positives (erroneous exclusion of safe compounds). As demonstrated in Figure 7, F1-score is more informative than accuracy alone when assessing classifier reliability under imbalanced conditions.

Furthermore, this comparison emphasizes the importance of pairing model depth with appropriate loss function design. While expressive models like MLP benefit from advanced losses such as Focal, even simpler networks can achieve strong performance if combined strategically. This insight offers a practical guideline for future applications targeting high-stakes biomedical classification tasks.

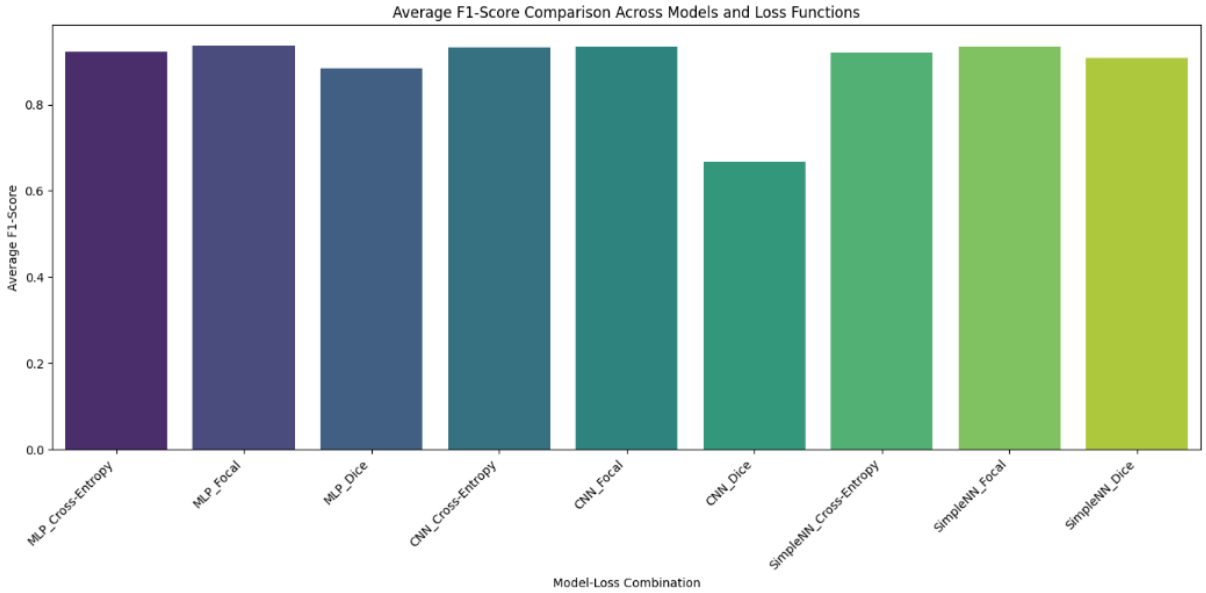

Figure 7. Average F1-Scores Across Model And Loss Function Combinations

3.1.6 Training and Validation Curves

To complement the evaluation metrics and improve insight into the training process, Figure 8 presents the average training and validation curves, both in terms of loss and accuracy, for each architecture and loss function. These plots help to illustrate convergence behavior and generalization capacity over multiple epochs.

For the MLP model, the training phase displayed smooth convergence, particularly when paired with Focal Loss. The validation accuracy remained in close alignment with training accuracy across epochs, reflecting strong generalization and minimal signs of overfitting. Dice Loss also demonstrated steady learning with stable accuracy patterns, although its validation performance plateaued earlier than Focal or Cross-Entropy.

For the CNN architecture, model learning showed moderate progression. Training with Binary Cross-Entropy resulted in relatively high training accuracy, but a noticeable divergence emerged between training and validation losses in later epochs, indicating mild overfitting. Focal Loss enabled smoother and more stable accuracy growth. Dice Loss showed minimal performance improvements and kept both training and validation loss curves relatively flat, suggesting learning stagnation.

In the 2-Layer Neural Network configuration, training behavior was less stable overall. Focal and Cross-Entropy Loss facilitated gradual accuracy increases, yet a visible gap remained between training and validation curves, especially under Dice Loss. This gap points to model overfitting and under-generalization, most likely due to the limited representational capacity of the shallow architecture.

Overall, the learning curves reinforce earlier metric-based insights: model architecture complexity and loss function selection have a substantial influence not only on final evaluation metrics but also on training dynamics. The use of Focal Loss, particularly in conjunction with deeper models such as MLP and CNN, helped maintain consistent validation trajectories and promoted smoother convergence behavior. Beyond visual inspection, these learning curves also highlight how the interplay between model complexity and loss function design influences convergence efficiency and gradient stability. Deeper architectures such as MLP and CNN exhibit more consistent descent in loss, implying stronger representational adaptability and more robust error minimization, especially under Focal Loss. This indicates that loss scaling not only mitigates class imbalance but also stabilizes gradient propagation during optimization. Conversely, the 2-Layer Neural Network tends to experience noisier updates, demonstrating its limited capacity to retain high-dimensional feature interactions. Such insights affirm that convergence smoothness is an indirect yet critical signal of both numerical stability and generalization resilience across epochs. Figure 8 illustrates the training and validation trends for loss and accuracy across all models, highlighting convergence behavior and generalization stability.

Copyright © 2025 **Author**, Page 1576
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

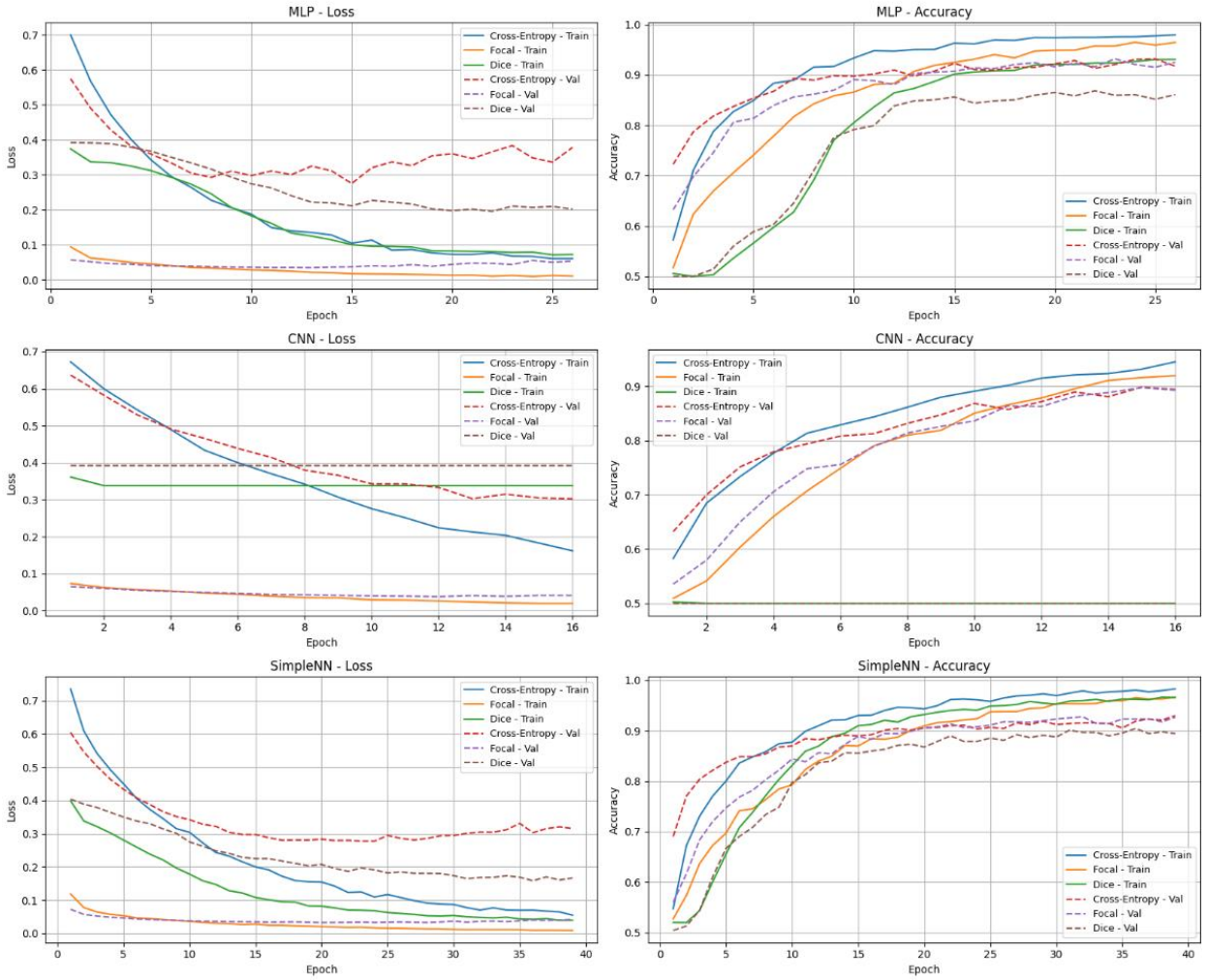

Figure 8. Trends of Training Vs. Validation Loss And Accuracy Per Model.

3.2 Discussion

This research offers a thorough analysis of the interplay between loss functions and model architecture in the context of predicting drug-induced autoimmunity based on molecular descriptors. Through multiple experimental setups and statistical validation, the study highlights how Focal Loss consistently delivers superior outcomes, particularly in challenging class-imbalanced scenarios.

Among all tested configurations, Focal Loss emerged as the most effective, showing high F1-scores and AUC values across different architectures. This can be attributed to its design, which assigns greater weight to misclassified or minority-class samples. As a result, models trained with this loss function are better at distinguishing immunogenic compounds that are typically underrepresented in clinical datasets. The advantage was especially pronounced when combined with deeper architectures like MLP, where the increased representational capacity further enhanced discriminatory power without sacrificing stability across validation splits.

On the other hand, Dice Loss, although popular in biomedical image segmentation tasks, did not perform well in this classification context. Particularly when paired with CNN, the model collapsed into predicting a single class. This behavior is indicative of convergence issues, potentially caused by the interaction between the Dice objective and the architecture's learning dynamics. In shallow models like 2-Layer Neural Network, Dice Loss also led to an increase in false positives and unstable learning curves, emphasizing its limited suitability for tabular classification unless augmented by architectural or regularization mechanisms.

Meanwhile, Binary Cross-Entropy (BCE) produced acceptable accuracy and precision but lagged behind in recall and F1-score. Its tendency to underperform under class imbalance reinforces previous findings in the literature, where BCE often fails to highlight minority class contributions unless complemented by sampling techniques or custom weighting strategies.

In terms of architecture, MLP demonstrated the most consistent and robust performance. Its multi-layer configuration, equipped with dropout regularization, enabled the model to learn complex feature interactions and generalize well across different validation folds. CNNs, although capable of capturing local feature interactions, were hindered by the absence of spatial structure in molecular descriptor data, making them more sensitive to

Copyright © 2025 **Author**, Page 1577
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

hyperparameter choices and initialization. 2-Layer Neural Network, with its limited depth, showed the least reliable performance, particularly under loss functions that demand strong learning capacity such as Dice.

Training stability analysis further reinforced these conclusions. MLP with Focal Loss consistently recorded the lowest standard deviations in metrics such as AUC and F1-score, confirming its suitability for clinical-grade prediction tasks. Conversely, 2-Layer Neural Network combined with Dice Loss yielded the highest variability, reflecting unreliable behavior in sensitive applications.

Complementing the metrics, learning curves demonstrated clear convergence trends. MLP and CNN models using Focal Loss showed closely aligned training and validation trajectories, suggesting minimal overfitting and effective generalization. 2-Layer Neural Network, in contrast, showed clear signs of overfitting, especially when trained under Dice Loss, where the validation accuracy stagnated early while training accuracy continued to rise.

Altogether, this comprehensive evaluation confirms that the pairing of a robust architecture with a well-suited loss function is essential for achieving dependable performance in high-stakes biomedical classification. The results underscore that performance must be assessed not only by scalar metrics, but also by consistency across folds, stability of training behavior, and interpretability of confusion matrices and ROC curves.

4. CONCLUSION

This study systematically compared three loss functions Binary Cross-Entropy (BCE), Focal Loss, and Dice Loss across three neural network architectures: Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and 2-Layer Neural Network. The results show that the MLP model optimized with Focal Loss achieved the best classification performance, reaching 94% accuracy, 93% precision, 95% recall, 94% F1-score, and an AUC of 0.97, while CNN and 2-Layer Neural Network models yielded lower but consistent outcomes under the same configurations. These results indicate that selecting appropriate loss functions and model depth is critical for handling class imbalance in biomedical data. The use of Focal Loss improved the model's sensitivity to the minority class, confirming its suitability for rare-event prediction such as drug-induced autoimmunity detection. Furthermore, SMOTE was applied only on the training folds during cross-validation, ensuring methodological validity and preventing data leakage. The contribution of this research to the development of science lies in providing an empirical benchmark for understanding how loss function selection influences deep learning performance in non-image biomedical data, particularly molecular descriptor-based prediction. This work bridges methodological gaps between computer vision-oriented loss functions and their adaptation to biomedical informatics, offering a reproducible framework for AI-driven pharmacovigilance and predictive toxicology. Future research may explore hybrid ensemble models and optimization techniques to further enhance predictive accuracy in biomedical applications.

REFERENCES

[1] D. Yang, X. Peng, S. Zheng, and S. Peng, "Deep learning-based prediction of autoimmune diseases," *Sci. Rep.*, vol. 15, no. 1, pp. 1–15, 2025, doi: 10.1038/s41598-025-88477-4.
[2] L. Huang, P. Liu, and X. Huang, "InterDIA: Interpretable prediction of drug-induced autoimmunity through ensemble machine learning approaches," *Toxicology*, vol. 511, p. 154064, Feb. 2025, doi: 10.1016/J.TOX.2025.154064.
[3] A. Begum and R. Kumar, "Design an Archetype to Predict the impact of diet and lifestyle interventions in autoimmune diseases using Deep Learning and Artificial Intelligence," 2022, [Online]. Available: https://www.researchsquare.com/article/rs-1405206/v1
[4] S. Singh, R. Kumar, S. Payra, and S. K. Singh, "Artificial Intelligence and Machine Learning in Pharmacological Research: Bridging the Gap Between Data and Drug Discovery.," *Cureus*, vol. 15, no. 8, p. e44359, Aug. 2023, doi: 10.7759/cureus.44359.
[5] UCI Machine Learning Repository, "Drug Induced Autoimmunity Prediction Data Set." University of California, Irvine, 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/1104/drug_induced_autoimmunity_prediction
[6] N. K. Iyortsuun, S. H. Kim, M. Jhon, H. J. Yang, and S. Pant, "A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis," *Healthc.*, vol. 11, no. 3, 2023, doi: 10.3390/healthcare11030285.
[7] S. Jadon, "A survey of loss functions for semantic segmentation," *2020 IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. CIBCB 2020*, 2020, doi: 10.1109/CIBCB48159.2020.9277638.
[8] E. Karakullukcu, "Leveraging convolutional neural networks for image-based classification of feature matrix data," *Expert Syst. Appl.*, vol. 281, p. 127625, Jul. 2025, doi: 10.1016/J.ESWA.2025.127625.
[9] R. Shwartz-Ziv, M. Goldblum, Y. L. Li, C. B. Bruss, and A. G. Wilson, "Simplifying Neural Network Training Under Class Imbalance," *Adv. Neural Inf. Process. Syst.*, vol. 36, no. NeurIPS, 2023.
[10] C. Wang, J. Balazs, G. Szarvas, P. Ernst, L. Poddar, and P. Danchenko, "Calibrating Imbalanced Classifiers with Focal Loss: An Empirical Study," *EMNLP 2022 - Proc. 2022 Conf. Empir. Methods Nat. Lang. Process. Ind. Track*, pp. 155–163, 2022, doi: 10.18653/v1/2022.emnlp-industry.14.
[11] P. Lin, Tsung-Yi and Goyal, Priya and Girshick, Ross and He, Kaiming and Dollár, "Focal Loss for Dense Object Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020, doi: 10.1109/TPAMI.2018.2858826.
[12] "An Introduction to Focal Loss | Baeldung on Computer Science." Accessed: Oct. 21, 2025. [Online]. Available: https://www.baeldung.com/cs/focal-loss
[13] X. Ma *et al.*, "On the value of imbalance loss functions in enhancing deep learning-based vulnerability detection," *Expert Syst. Appl.*, vol. 291, p. 128504, Oct. 2025, doi: 10.1016/J.ESWA.2025.128504.
[14] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li, "Dice loss for data-imbalanced NLP tasks," *Proc. Annu. Meet. Assoc.*

Copyright © 2025 **Author**, Page 1578
This Journal is licensed under a Creative Commons Attribution 4.0 International License

Building of Informatics, Technology and Science (BITS)
Volume 7, No 3, December 2025 Page: 1568−1579
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i3.8581

Comput. Linguist., no. 2, pp. 465–476, 2020, doi: 10.18653/v1/2020.acl-main.45.

[15] H. Fan, W. Yan, L. Wang, J. Liu, Y. Bin, and J. Xia, "Deep learning-based multi-functional therapeutic peptides prediction with a multi-label focal dice loss function," *Bioinformatics*, vol. 39, no. 6, p. btad334, Jun. 2023, doi: 10.1093/bioinformatics/btad334.

[16] M. Yeung, E. Sala, C. B. Schönlieb, and L. Rundo, "Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Comput. Med. Imaging Graph.*, vol. 95, no. December 2021, 2022, doi: 10.1016/j.compmedimag.2021.102026.

[17] T. Li, H. Fan, J. Zhao, X. Yang, and J. Xia, "MultiPep-DLCL: recognition of multifunctional therapeutic peptides through deep learning with label-sequence contrastive learning.," *Brief. Bioinform.*, vol. 26, no. 3, May 2025, doi: 10.1093/bib/bbaf274.

[18] M. Cabezas and Y. Diez, "An Analysis of Loss Functions for Heavily Imbalanced Lesion Segmentation," *Sensors*, vol. 24, no. 6, pp. 1–21, 2024, doi: 10.3390/s24061981.

[19] "Deep Learning for Tabular Data | Medium," Strong Analytics. Accessed: Oct. 21, 2025. [Online]. Available: https://medium.com/@stronganalytics/deep-learning-for-tabular-data-an-overview-3a742d72246b

[20] M. Li, X. Zhang, C. Thrampoulidis, J. Chen, and S. Oymak, "AutoBalance: Optimized Loss Functions for Imbalanced Data," *Adv. Neural Inf. Process. Syst.*, vol. 4, no. NeurIPS, pp. 3163–3177, 2021.

[21] Z. Małyjurek, D. de Beer, E. Joubert, and B. Walczak, "Combining class-modelling and discriminant methods for improvement of products authentication," *Chemom. Intell. Lab. Syst.*, vol. 228, p. 104620, Sep. 2022, doi: 10.1016/J.CHEMOLAB.2022.104620.

[22] X. Liu, L. Wang, L. Ma, and C. Wang, "DRFL: Dynamic-Recall Focal Loss for Image Classification and Segmentation," *Appl. Artif. Intell.*, vol. 38, no. 1, 2024, doi: 10.1080/08839514.2024.2411845.

[23] M. N. Razali, N. Arbaiy, P. C. Lin, and S. Ismail, "Optimizing Multiclass Classification Using Convolutional Neural Networks with Class Weights and Early Stopping for Imbalanced Datasets," *Electron.*, vol. 14, no. 4, pp. 1–14, 2025, doi: 10.3390/electronics14040705.

[24] S. Rajaraman, P. Ganesan, and S. Antani, "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks," *PLoS One*, vol. 17, no. 1, p. e0262838, Jan. 2022, doi: 10.1371/journal.pone.0262838.

[25] M. Yeung, L. Rundo, Y. Nan, E. Sala, C. B. Schönlieb, and G. Yang, "Calibrating the Dice Loss to Handle Neural Network Overconfidence for Biomedical Image Segmentation," *J. Digit. Imaging*, vol. 36, no. 2, pp. 739–752, 2023, doi: 10.1007/s10278-022-00735-3.

[26] J. He, Z. Liu, and X. Tang, "A deep learning model for predicting systemic lupus erythematosus-associated epitopes," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, 2025, doi: 10.1186/s12911-025-03056-x.

[27] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.

[28] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, p. 100994, 2024, doi: 10.1016/j.patter.2024.100994.

This Journal is licensed under a Creative Commons Attribution 4.0 International License