

Analysis of the Impact of Dimensionality Reduction on Deep Learning Performance for the Classification of Microarray and RNA-Seq Gene Expression Data

Duwi Lufita Marfiana^{1*}, Muhammad Haris²

^{1,2}Computer Science Department, Faculty of Information Technology, Universitas Nusa Mandiri, DKI Jakarta, Indonesia

E-mail: ¹*14230015@nusamandiri.ac.id, ²muhammad.uhs@nusamandiri.ac.id

(Received: 21 Oct 2025, revised: 19 Dec 2025, accepted: 8 Jan 2026)

Abstract

One of the most popular and widely used methods in various fields, not only in information technology, including bioinformatics. With the increasing complexity of data and problems faced, ML offers solutions for analyzing and interpreting very large and dense data. This study aims to analyze gene expression, which has a crucial role in disease diagnosis, especially cancer, requiring techniques capable of identifying patterns and relationships in complex genetic data. Several techniques that play a role in improving the performance of deep learning models, including PCA reduction techniques, TruncatedSVD, PaCMAP, and TorchDR, successfully simplified data with high dimensionality without eliminating important information structures, thereby speeding up the training process and improving prediction accuracy. DNN was able to achieve a perfect score with accuracy and F1-Score 1 and remained superior to TorchDR with accuracy using test data of 0.9990, PaCMAP 0.979, and TruncatedSVD 0.988. DNN is able to obtain a perfect score with accuracy and F1-Score 1, and remains superior to TorchDR with accuracy using test data of 0.9990, PaCMAP 0.979 and TruncatedSVD 0.988. The dimensionality reduction techniques of PaCMAP and TorchDR provide a more stable data representation, by producing a more separated data distribution and supporting high performance in DNN and MLP models, which makes it the right choice in the dimensionality reduction method of gene expression data.

Keywords: Dimension Reduction, Deep Learning, Gene Expression, RNA-Seq, Bioinformatics

I. INTRODUCTION

In today's digital era, machine learning (ML) has become one of the most popular and widely used methods in various fields, not only in information technology, but also in bioinformatics. With the increasing complexity of data and the problems faced, ML offers a solution for analyzing and interpreting very large and dense data sets. In this paper, in the context of bioinformatics, the ever-growing data size, particularly in gene expression and microarrays, is driving researchers to shift from traditional methods to machine learning-based approaches. Gene expression analysis, which plays a crucial role in disease diagnosis, particularly cancer, requires techniques capable of identifying patterns and relationships in complex genetic data [1].

Gene expression data, particularly those obtained from microarray and RNA-Seq technologies, allow the simultaneous analysis of thousands of genes, providing an in-depth picture of the genetic profile of cancer cells [2]. The

use of microarray data in cancer classification has become a major focus in bioinformatics research. Microarray data allows the processing of thousands of gene analyses, providing a detailed picture of the gene expression profile in cancer cells. Machine learning-based classification methods, such as Support Vector Machine (SVM) and Random Forest, have been successfully applied to analyze microarray data in previous studies [3]. Microarrays allow the simultaneous measurement of thousands of genes in a single experiment. In the context of cancer diagnosis, microarray data provide in-depth insights into the differential gene expression profiles between cancer cells and normal cells, making them invaluable tools for cancer classification and identification.

To address these challenges, the use of deep learning has emerged as a viable approach [4], [5]. Deep learning methods are capable of rapidly extracting complex features from genetic data, enhancing the ability to identify patterns that may be invisible with traditional methods. By integrating effective feature selection techniques, such as Information



Gain, deep learning is expected to improve the accuracy of microarray data classification, making it a highly effective tool in supporting more precise and rapid cancer diagnosis.

To address these challenges, Deep Learning (DL)-based approaches have demonstrated superior performance in cancer classification based on gene expression data. DL is capable of automatically extracting non-linear and complex feature representations from raw data without the need for manual feature extraction. Architectures such as Multilayer Perceptron (MLP), Deep Neural Network (DNN), and TabNet have been widely used in cancer research and proven effective [5], [6], [7].

In a recent study, Gupta et al. (2022) evaluated various DL models on microarray gene expression data for cancer classification. They demonstrated that the use of appropriate optimization and feature selection techniques can significantly improve classification accuracy, even in highly imbalanced data [6]. Another study by Zeng et al. (2025) highlighted the need for a comprehensive approach to address the combined challenges of high dimensionality, under sampling, and multi-class imbalance in gene expression data. They proposed the MGAN-FB (Multi-classification Generative Adversarial Network with Feature Bundling) model, which integrates feature bundling, autoencoder, and multi-class discriminator techniques to produce more accurate and stable cancer classification [7].

This study focuses on overcoming the challenges posed by high-dimensional gene expression data by integrating dimension reduction techniques with deep learning models. The proposed approach evaluates several dimension reduction methods combined with deep learning architectures such as DNN, MLP, TabNet, and SAN to improve classification accuracy while preserving biologically relevant features. Through systematic experiments, this study aims to analyze the performance, stability, and generalization of models under different feature representations [8], [9].

missing values and feature normalization. After preprocessing, dimension reduction techniques are applied to transform high-dimensional gene expression data into lower-dimensional representations.

The reduced data is then divided into training and testing sets, followed by class imbalance analysis and balancing. Several deep learning architectures are then trained and evaluated using cross-validation strategies. Model performance is assessed using quantitative metrics, including accuracy, precision, recall, F1-score, confusion matrix, and ROC curve. This structured procedure ensures a systematic evaluation of the impact of dimension reduction techniques on deep learning performance.

1. Dataset Collection

The dataset used in this study consists of RNA-Seq gene expression data covering 801 samples from five types of cancer: BRCA, COAD, KIRC, LUAD, and PRAD. Each sample represents gene expression measurements across thousands of genes, resulting in a high-dimensional feature space. This dataset includes multi-class labels and exhibits class imbalance, making it suitable for evaluating the effectiveness of dimension reduction and deep learning techniques in complex bioinformatics scenarios [8].

2. Data Preprocessing

a. Handling missing value

Handling missing values is performed as an initial preprocessing step to ensure data quality and computational stability. In this study, rows containing missing values were removed from the dataset. This strategy was chosen because the proportion of missing values was relatively small and randomly distributed across the sample, minimizing the risk of significant information loss or class bias.

Removing incomplete records helps prevent distortion in dimension reduction and deep learning training processes, which are sensitive to undefined or zero values. After the deletion process, the dataset retains a sufficient sample in each class to support reliable training and evaluation.

b. Normalization

Feature normalization is applied as a pre-processing step to rescale gene expression values before dimensionality reduction and deep learning model training. In this study, Min-Max normalization is used to transform each feature into a standard range between 0 and 10, using the following Formula 1.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where x represents the original gene expression value, and x_{min} and x_{max} indicate the minimum and maximum values of each feature, respectively.

Min-Max normalization was chosen because it preserves the original data distribution while ensuring that all features contribute proportionally during model training. This characteristic is particularly important for gene expression

II. METHODS

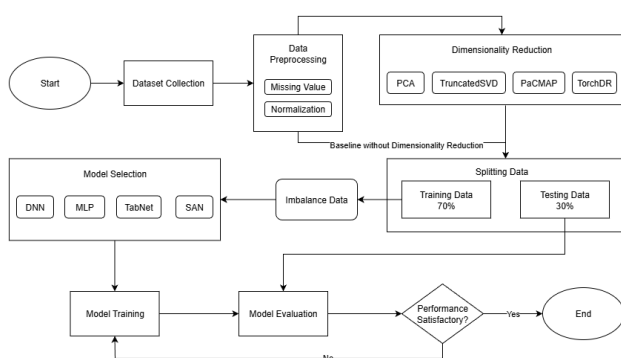


Figure 1. Flowchart of the Proposed Research Methodology.

Figure 1 illustrates the overall research workflow used in this study. The methodology begins with dataset collection, followed by data preprocessing, which includes handling

data, where feature magnitudes can vary significantly across genes and, if not normalized, can dominate distance-based and gradient-based learning processes.

The impact of normalization on data distribution is illustrated in Figure 2, which shows that normalized gene expression values exhibit a more uniform feature scale compared to the original data. By reducing scale differences between features, normalization improves numerical stability during training and supports more effective dimensionality reduction and model convergence.

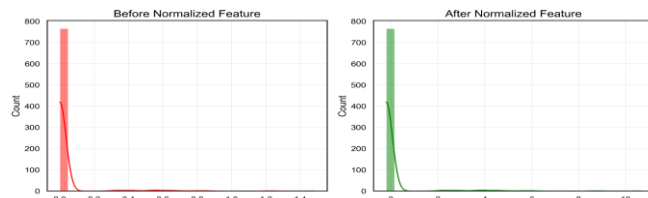


Figure 2. The Results of the Normalization of Gene Expression Values After Normalization to Ensure a Uniform Feature Scale

3. Dimensionality Reduction

In this study, the reduced number of dimensions was kept constant across all dimension reduction techniques to ensure a fair comparison. For PCA and TruncatedSVD, the number of components was chosen to retain the most informative features while significantly reducing the dimensions, following common practice in gene expression analysis.

PaCMAP was applied using the default environment and pair ratio parameters, as recommended in previous studies, to preserve local and global data structures. TorchDR was used as the framework for nonlinear embedding, using PCA-based and t-SNE transformations provided by the library. This configuration was chosen to evaluate the effectiveness of linear and nonlinear dimension reduction techniques under consistent experimental conditions.

4. Splitting Data

Data splitting is performed after preprocessing and dimensionality reduction to avoid data leakage. The dataset is divided into a training set (70%) and a test set (30%). Class imbalance analysis and balancing are only applied to the training data, while the test set is left unchanged to maintain its role as an unbiased evaluation set. This sequence ensures that information from the test data does not influence model training or parameter tuning, thereby maintaining the validity of the evaluation results.

5. Imbalance Data

Gene expression datasets often exhibit class imbalance, which can bias the learning process and reduce classification performance, especially for minority classes (Figure 3). To address this issue, this study applies a random oversampling strategy to balance class distribution during model training [9].

Oversampling is performed exclusively on the training data, both in cross-validation folds and final model training,

while the validation and testing sets remain unchanged. This approach prevents information leakage and ensures that performance evaluation reflects the model's true generalization ability.

By balancing the training data (Figure 4), the model is encouraged to learn representative patterns from all classes, thereby improving robustness and yielding more reliable evaluation results across all accuracy, precision, recall, and F1-score metrics.

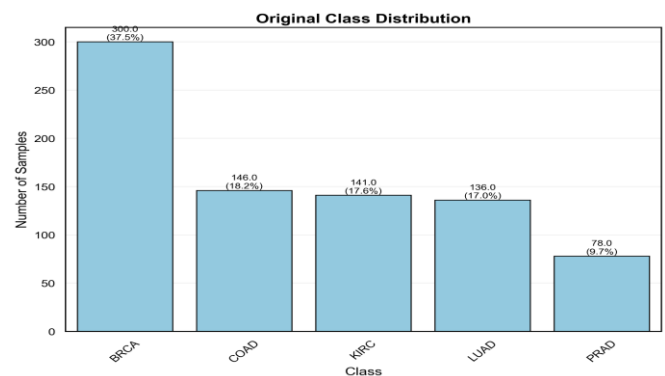


Figure 3. Original Class Distribution Results

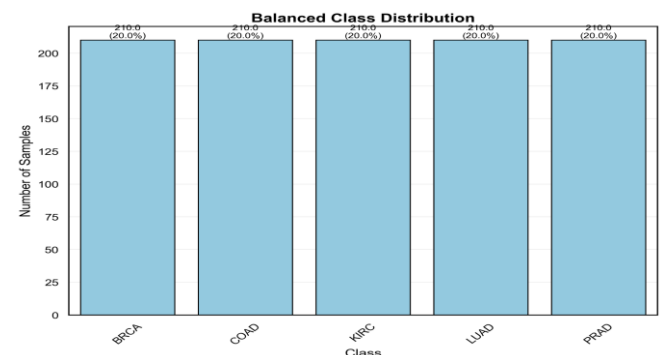


Figure 4. Balance Class Distribution Results

6. Model Selection

In this study, several deep learning architectures were selected to evaluate their effectiveness in classifying high-dimensional gene expression data after dimensionality reduction, namely Deep Neural Network (DNN), Multilayer Perceptron (MLP), TabNet, and Self-Attention Network (SAN). DNN and MLP were selected [10], [11], which demonstrated that fully connected deep learning architectures are effective for cancer classification using high-dimensional microarray and RNA-Seq gene expression data. Their ability to model complex non-linear relationships makes DNN and MLP suitable baselines for evaluating classification performance after dimensionality reduction. TabNet and SAN are included to evaluate attention-based deep learning architectures on low-dimensional gene expression data. TabNet enables the utilization of selective features for tabular inputs, while SAN focuses on modeling feature relationships through self-attention. The inclusion of both allows analysis of how attention mechanisms respond to

various dimension reduction techniques without relying on domain-specific architecture assumptions.

Table 1. Configuration Details of Deep Learning Models Evaluated in This Study

Aspect	DNN	MLP	TabNet	SAN
Hidden Layers / Blocks	2 hidden layers	3 hidden layers	5 decision steps	Attention + feed-forward blocks
Hidden Units	128-64	256-128-64	n_d = 64, n_a = 64	Hidden dimension = 128
Activation Function	ReLU	ReLU	ReLU	ReLU [12]
Optimizer	Adam	Adam	Adam (lr=0.02)	Adam (lr=0.001)
Batch Size	32	32	256 (vBatch 128)	32

Table 1 summarizes the architectural configurations of the evaluated models, which were kept consistent to ensure fair comparison and to focus on the impact of dimension reduction.

7. Model Training

The model training process was conducted using a 10-fold cross-validation method, with each fold having 200 epochs to measure accuracy, precision, recall, and F1 scores.

8. Model Evaluation

In the final stage, the model was evaluated using Training Accuracy Across Folds, Validation Accuracy on Test Set, and Confusion Matrix to evaluate model performance during training and testing on the split data. After model evaluation, performance results will be assessed to determine whether the model achieves satisfactory performance. If the result is no, the training process is repeated with adjusted training configurations, such as hyperparameters or training settings. This iterative training-evaluation cycle continues until acceptable performance is achieved.

III. RESULTS AND DISCUSSION

1) Baseline Performance Without Dimensionality Reduction

A baseline evaluation was conducted to assess the classification performance of the selected deep learning models such as DNN, MLP, TabNet, and SAN using normalized gene expression data without applying any dimension reduction techniques. This baseline evaluation serves as a reference for understanding how each model performs in the original high-dimensional feature space and for measuring the impact of dimension reduction methods in subsequent experiments.

As shown in Table 2, DNN achieved the highest baseline performance, with precision, recall, F1-score, and accuracy values of 0.99. These results indicate that DNN is capable of

effectively learning complex non-linear patterns directly from high-dimensional gene expression data. MLP also shows strong performance, with an accuracy of 0.98 and an F1-score of 0.97, confirming its robustness as a fully connected neural network architecture for gene expression classification tasks.

Table 2. Baseline Classification Performance Without Dimensionality Reduction

Model	Precision	Recall	F1-Score	Accuracy
DNN	0.99	0.99	0.99	0.99
MLP	0.98	0.97	0.97	0.98
TabNet	0.91	0.92	0.91	0.93
SAN	0.32	0.34	0.29	0.28

TabNet achieves a competitive baseline accuracy of 0.93 and an F1-score of 0.91. Although its performance is slightly lower than that of DNN and MLP, TabNet is still able to maintain reliable classification results in most classes. A small decrease in performance is mainly observed in cancer types such as KIRC and LUAD, as reflected in the confusion matrix (Figure 5) and ROC curve (Figure 6), suggesting that TabNet may require more structured or concise feature representations to fully utilize its attention-based feature selection mechanism.

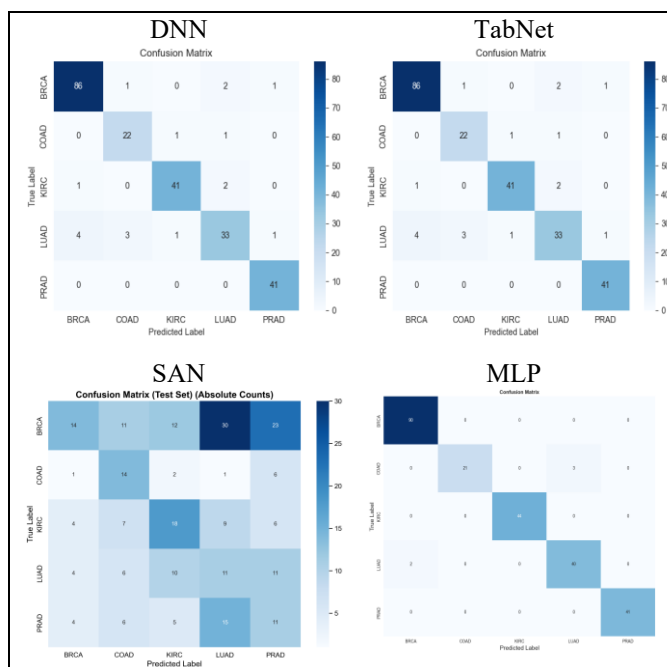


Figure 5. Confusion Matrices of Baseline Classification Results Without Dimensionality Reduction

In contrast, SAN showed the lowest baseline performance among all models, with an accuracy of 0.28 and an F1 score of 0.29. The confusion matrix in Figure 5 reveals a highly uneven distribution of predictions, where the model failed to consistently distinguish between several cancer classes. This behavior is further supported by the ROC curve in Figure 6, which shows an unstable curve and low AUC

values for several classes, with the lowest AUC observed for LUAD (around 0.49). These results indicate that SAN struggles to learn meaningful feature relationships directly from high-dimensional gene expression data without dimension reduction [13].

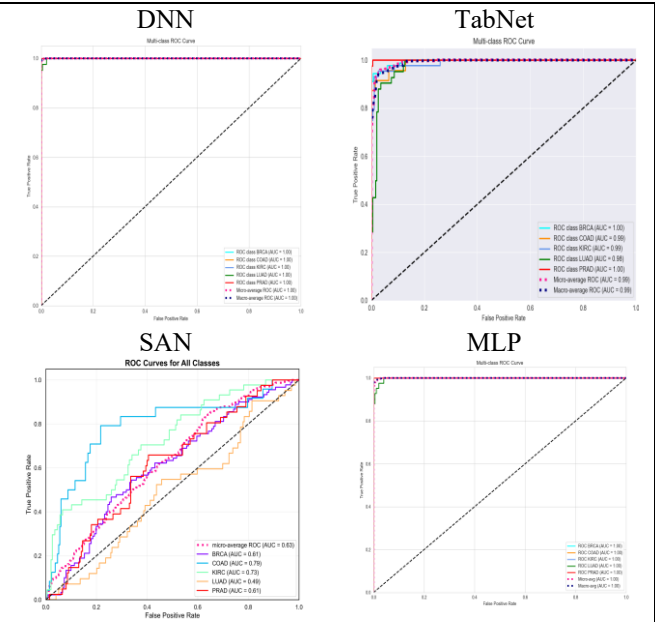


Figure 6. ROC Curve of Baseline Classification Results Without Dimensionality Reduction

The preliminary results indicate that while fully connected architectures such as DNNs and MLPs can achieve strong performance in the native feature space, attention-based models particularly SAN are highly sensitive to feature dimensions and data structure. These findings highlight the need to apply dimension reduction techniques to improve feature representation, enhance class separability, and support more stable performance across different deep learning architectures in future experiments.

2) PCA

After establishin baseline performance, Principal Component Analysis (PCA) was applied as a linear dimension reduction technique to transform the original high-dimensional gene expression data into a lower-dimensional space while retaining maximum variance. The objective of this experiment was to evaluate how linear dimension reduction affects class separability and the performance of various deep learning models.

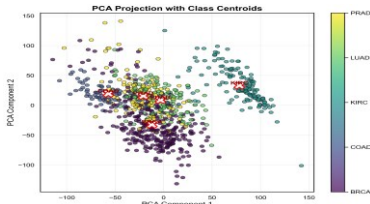


Figure 7. Projection of the Midpoint of the Class Distribution in the PCA Reduction Technique

As illustrated in Figure 7, the PCA projection shows that some cancer classes, such as KIRC, form relatively distinct clusters, while other classes show some overlap in the reduced two-dimensional space. This overlap indicates that although PCA is effective in capturing global variance, it may not fully preserve the non-linear class boundaries inherent in the gene expression data.

Table 3. Classification of Test Data Results

Model	PCA Precision	PCA Recall	PCA F1 Score	PCA Accuracy
DNN	1	0.99	0.99	0.99
MLP	0.98	0.97	0.97	0.98
TabNet	0.94	0.93	0.93	0.94
SAN	0.51	0.56	0.5	0.49

The classification results using PCA are summarized in Table 3. DNN achieved the best overall performance, with an accuracy of 0.99 and an F1 score of 0.99, indicating that DNN is highly capable of utilizing PCA-transformed features for accurate classification. MLP also performed competitively, achieving an accuracy of 0.98 and an F1 score of 0.97, confirming its robustness under linear dimension reduction.

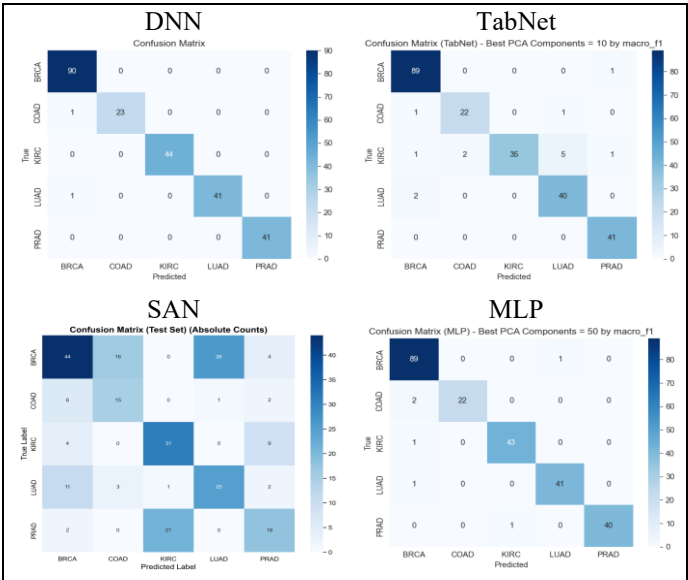


Figure 8. Confusion Matrices of Classification Results Using PCA Dimensionality Reduction

TabNet shows moderate performance below PCA, with an accuracy of 0.94 and an F1 score of 0.93. Although these values remain above 0.90, the lower recall and F1 score compared to DNN and MLP indicate that TabNet is less effective at capturing discriminative information from PCA-reduced features. This behavior suggests that TabNet may benefit more from feature representations that preserve local or non-linear structures than from variance-based projections alone.

SAN shows significantly lower performance compared to other models, with an accuracy of 0.49 and an F1 score of

0.50. The confusion matrix presented in Figure 8 reveals frequent classification errors across various classes, reflecting SAN's difficulty in learning meaningful feature relationships [14] from PCA-transformed data. This observation is further supported by the ROC curve in Figure 9, which shows low and unstable AUC values for several classes, indicating weak discriminative power.

As shown in Figure 10, the TruncatedSVD projection reveals improved class separation for certain cancer types, such as KIRC, while some overlap remains between other classes. This distribution indicates that TruncatedSVD can preserve important global structures in the data, even though some non-linear relationships between classes may still be lost in the reduced space.

The classification results obtained using TruncatedSVD are summarized in Table 4. DNN again achieved the best overall performance, with an accuracy of 0.99 and an F1 score of 0.99, demonstrating strong and consistent classification capabilities on SVD-transformed features. MLP followed closely, with an accuracy of 0.97 and an F1 score of 0.96, confirming its robustness under linear dimension reduction.

Table 4. Classification of Test Data Results

Model	SVD Precision	SVD Recall	SVD F1-Score	SVD Accuracy
DNN	1	0.99	0.99	0.99
MLP	0.97	0.96	0.96	0.97
TabNet	0.96	0.95	0.95	0.95
SAN	0.55	0.59	0.55	0.54

Figure 9. ROC Curve of Baseline Classification Results Using PCA Dimensionality Reduction

PCA-based results show that linear dimension reduction can significantly improve performance compared to the baseline for fully connected architectures such as DNNs and MLPs. However, PCA appears to be less suitable for attention-based models, particularly SANs, which struggle to adapt to linear projections that do not preserve complex feature interactions. These findings suggest that while PCA is effective for reducing dimensions and improving computational efficiency, more advanced non-linear dimension reduction techniques may be necessary to support stable performance across a wider range of deep learning architectures.

3) TruncatedSVD

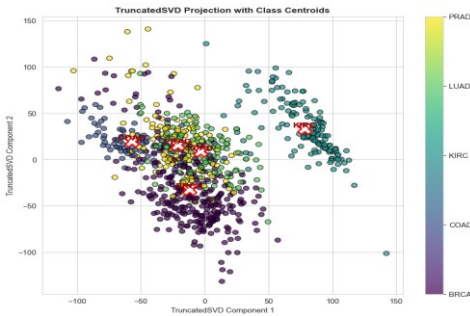


Figure 10. Projection of the Midpoint of the Class Distribution in the TruncatedSVD Reduction Technique

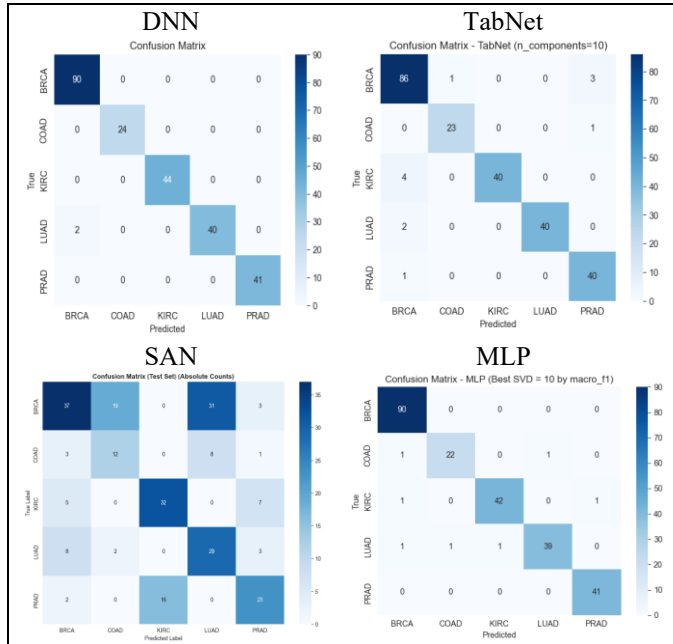


Figure 11. Confusion Matrices of Classification Results Using TruncatedSVD Dimensionality Reduction

In contrast, SAN continues to show significantly lower performance, with an accuracy of 0.54 and an F1 score of 0.55. The confusion matrix in Figure 11 reveals frequent

classification errors in several classes, indicating that SAN struggles to effectively utilize the features generated by TruncatedSVD. This behavior is further supported by the ROC curve in Figure 12, where the AUC value remains relatively low and unstable, despite a marginal improvement compared to the PCA results.

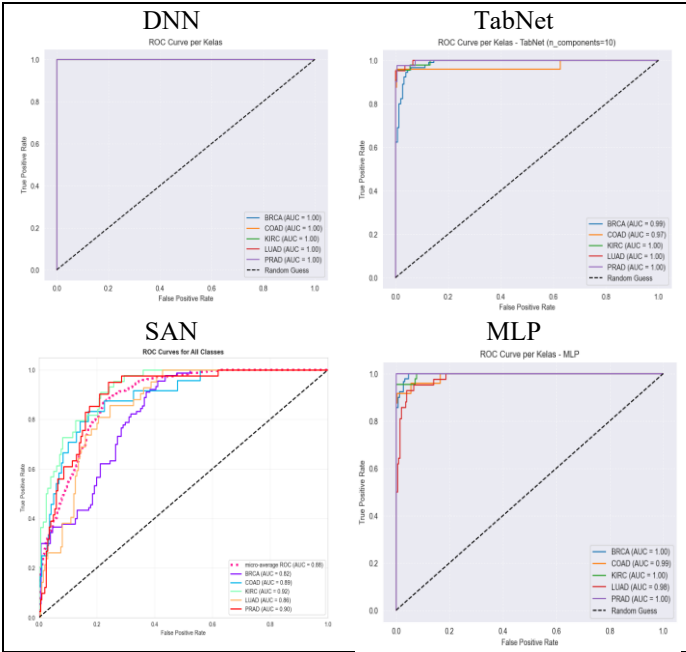


Figure 12. ROC Curve of Baseline Classification Results Using TruncatedSVD Dimensionality Reduction

TruncatedSVD results show a moderate performance improvement over PCA for some models, particularly TabNet, while maintaining strong performance for DNN and MLP. However, like PCA, TruncatedSVD remains less effective for attention-based architectures such as SAN. These findings suggest that although TruncatedSVD offers advantages in handling high-dimensional data, linear dimension reduction techniques alone may not be sufficient to support stable and robust performance across all deep learning models, motivating the exploration of non-linear dimension reduction methods in future experiments.

4) PacMAP

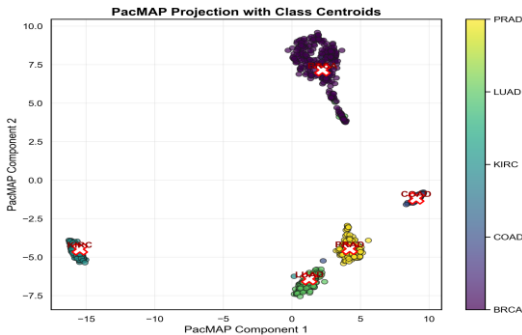


Figure 13 Projection of the Midpoint of the Class Distribution Using the PacMAP Reduction Technique

As illustrated in Figure 13, the PacMAP projection shows a well-separated and compact distribution of the five cancer classes (BRCA, COAD, KIRC, LUAD, and PRAD). Compared to PCA and TruncatedSVD, the overlap between classes is significantly reduced, indicating that PacMAP is more effective at preserving the intrinsic data structure and improving visual class separation in lower-dimensional space.

The classification results using PacMAP are summarized in Table 5. All models DNN, MLP, TabNet, and SAN achieved consistently high performance, with precision, recall, F1-score, and accuracy values ranging from 0.98 to 0.99. DNN, MLP, and TabNet each achieved an accuracy of 0.98 with an F1-score of 0.98, while SAN achieved a slightly higher accuracy of 0.9834. These results show that PacMAP provides stable and informative feature representations that support reliable classification across various deep learning architectures.

Table 5. Classification of Test Data Results

Model	PacMap Precision	PacMap Recall	PacMap F1 Score	PacMap Accuracy
DNN	0.99	0.98	0.98	0.98
MLP	0.99	0.98	0.98	0.98
TabNet	0.99	0.98	0.98	0.98
SAN	0.99	0.98	0.98	0.9834

The confusion matrix presented in Figure 14 further confirms the effectiveness of PacMAP. All models show a highly concentrated diagonal pattern, reflecting a high number of correct predictions across all cancer classes. Classification errors are minimal and evenly distributed, indicating that PacMAP effectively reduces feature redundancy and improves class discrimination.

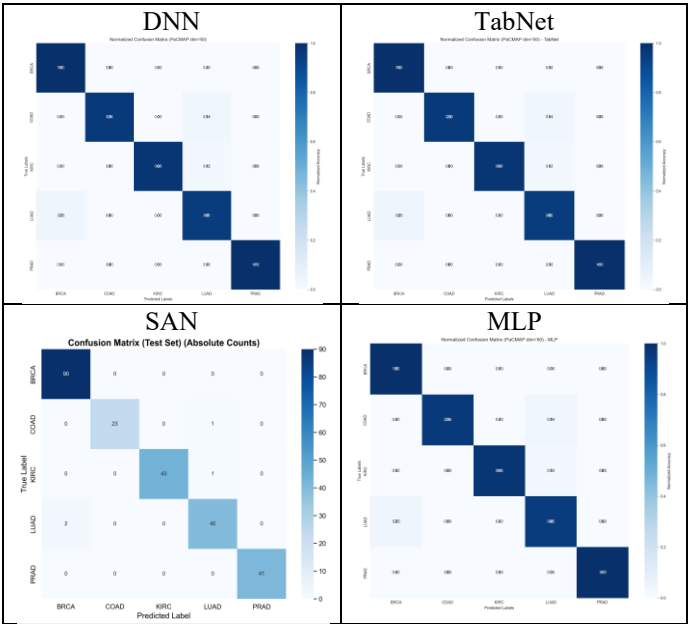


Figure 14. Confusion Matrices of Classification Results Using PacMAP Dimensionality Reduction

This observation is reinforced by the ROC curves shown in Figure 15, where all models exhibit smooth, wide curves with high AUC values across all classes. Consistently high AUC values indicate strong discriminatory power and stable performance, regardless of the underlying model architecture.

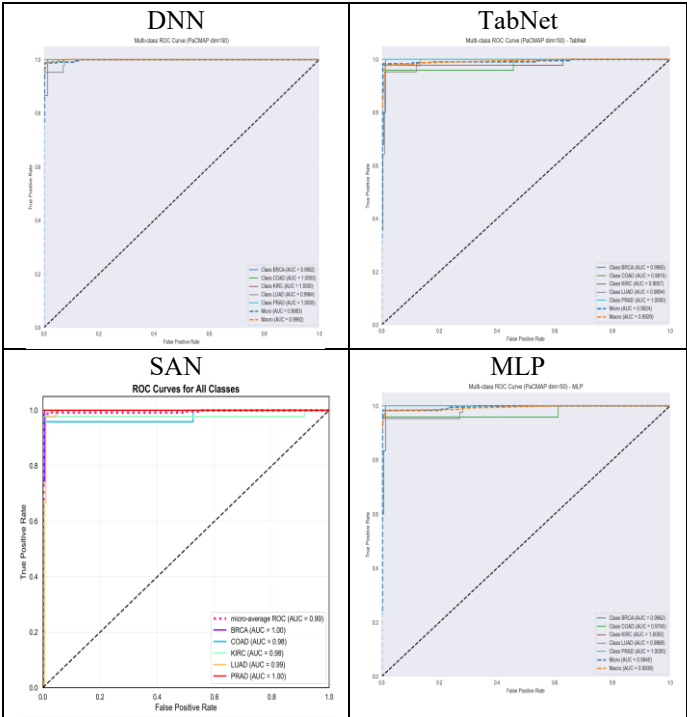


Figure 15. ROC Curve of Baseline Classification Results Using PaCMAP Dimensionality Reduction

After all, PaCMAP results show a significant improvement in performance compared to linear and basic dimension reduction techniques. Unlike PCA and TruncatedSVD, PaCMAP successfully supports attention-based models such as SAN, which previously showed unstable performance. These findings indicate that PaCMAP is highly effective in generating low-dimensional representations that retain important biological information, enabling consistent and robust classification performance across various deep learning models.

5) TorchDR

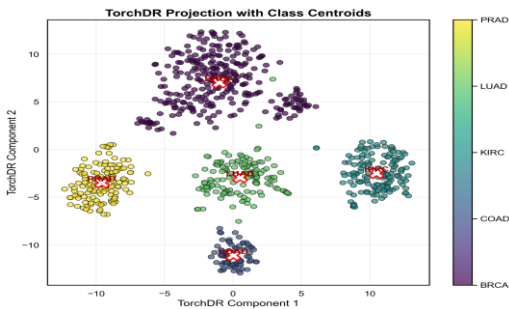


Figure 16. Projection of the Midpoint of the Class Distribution Using the Torchdr Reduction Technique

As summarized in Table 6, TorchDR produced the highest overall performance among all evaluated models. DNN achieved perfect classification results, with accuracy and F1-score values of 1.00, indicating complete class separation in the TorchDR. MLP and TabNet also showed strong performance, with accuracy values of 0.97 and 0.96 and F1-scores of 0.96 and 0.95, respectively. These results confirm that TorchDR provides highly discriminative feature representations that can be effectively utilized by fully connected and attention-based architectures.

The confusion matrix shown in Figure 16 reveals near-perfect diagonal dominance for all models, particularly for DNN, where no classification errors were observed across all five cancer classes. Minor classification errors in MLP and TabNet are limited and do not significantly affect overall performance, demonstrating strong generalization capabilities in the reduced feature space.

Table 6. Classification of Test Data Results

Model	TorchDR Precision	TorchDR Recall	TorchDR F1 Score	TorchDR Accuracy
DNN	1	1	1	1
MLP	0.97	0.96	0.96	0.97
TabNet	0.96	0.95	0.95	0.96
SAN	0.54	0.59	0.52	0.51

The ROC curves presented in Figure 17 further reinforce these findings. All models exhibit smooth ROC curves with AUC values close to or equal to 1.00 across all classes, indicating exceptional discriminatory power. Compared to previous dimension reduction techniques, TorchDR consistently produces the most stable and robust ROC behavior, with minimal performance variation between models.

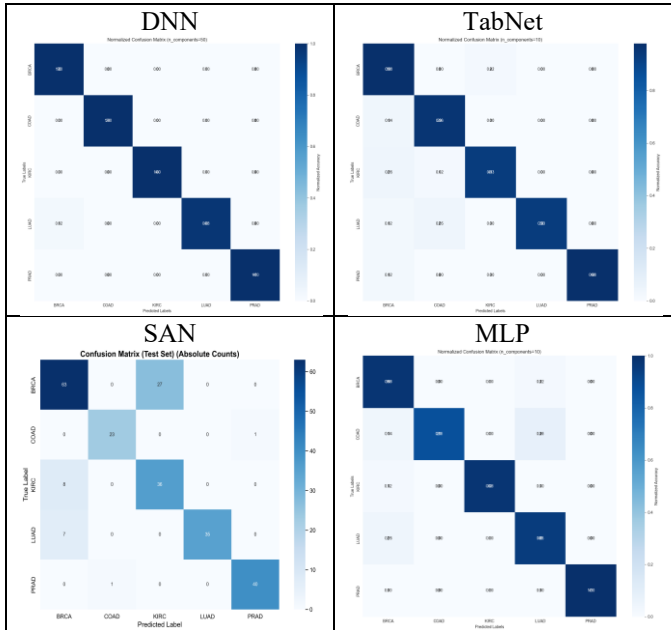


Figure 17. Confusion Matrices of Classification Results Using Torchdr Dimensionality Reduction

TorchDR's results represent the most effective dimension reduction results in this study (Figure 18). Compared to PaCMAP, TorchDR not only maintains high class separation but also improves model robustness, particularly for DNNs, which achieve perfect classification performance. These findings demonstrate that TorchDR is highly capable of capturing complex nonlinear structures in gene expression data, enabling superior and consistent classification performance across various deep learning architectures. The robust results obtained using TorchDR highlight its potential as a powerful dimension reduction technique for high-dimensional genomic data analysis.

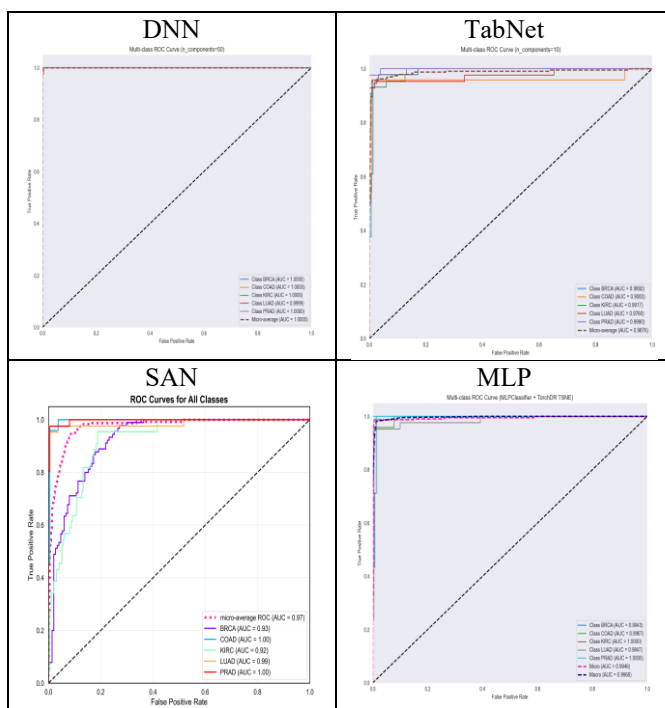


Figure 18. ROC Curve of Baseline Classification Results Using TorchDR Dimensionality Reduction

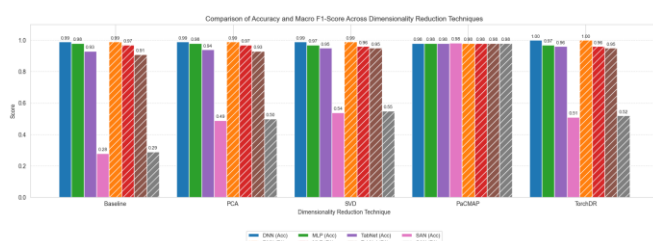


Figure 19. Comparison of Accuracy and Macro F1-Score Across Different Dimensionality Reduction Techniques

Figure 19 presents an overall comparison of model performance across all dimension reduction techniques evaluated using F1-Score and accuracy as the main metrics. The results show that dimension reduction consistently improves model performance compared to the baseline, demonstrating its effectiveness in handling high-dimensional gene expression data.

Linear reduction methods such as PCA and TruncatedSVD offer moderate and stable improvements, while non-linear approaches show a more pronounced impact. Specifically, PaCMAP produces the most consistent performance improvement across models, particularly for SAN architectures, highlighting its ability to preserve meaningful data structure. Overall, the figure confirms that the combination of dimension reduction and appropriate model architecture is crucial for achieving strong and generalizable classification performance.

IV. CONCLUSION

This study shows that dimensionality reduction plays an important role in improving the performance and stability of deep learning models for high-dimensional gene expression data. The experimental results show that the application of appropriate dimensionality reduction techniques consistently improves classification performance compared to the baseline scenario without reduction.

Among the evaluated methods, PaCMAP shows a significant impact, especially when combined with attention-based models such as SAN, resulting in more balanced improvements in both F1-Scores and accuracy. This suggests that preserving local and global data structures is beneficial for capturing complex feature relationships in gene expression datasets.

Overall, these findings confirm that integrating dimension reduction with deep learning architectures improves model generalization and robustness [13], [14], [15] highlighting the importance of selecting appropriate reduction techniques when dealing with high-dimensional bioinformatics data.

REFERENCES

- [1] B. Bagiroz, E. Doruk, and O. Yildiz, "Machine learning in bioinformatics: Gene expression and microarray studies," in Proc. 2020 Medical Technologies Congress (TIPTEKNO), 2020, doi: 10.1109/TIPTEKNO50054.2020.9299285.
- [2] N. Tabassum, M. A. S. Kamal, M. A. H. Akhand, and K. Yamada, "Cancer classification from gene expression using ensemble learning with an influential feature selection technique," BioMedInformatics, vol. 4, no. 2, pp. 1275–1288, 2024, doi: 10.3390/biomedinformatics4020070.
- [3] X. Deng and Y. Xu, "Cancer classification using microarray data by DPCAForest," in Proc. IEEE 31st Int. Conf. Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 2019, pp. 1081–1087, doi: 10.1109/ICTAI.2019.00151.
- [4] H. Chebli, Z. Mashhadieh, M. A. Ali, M. K. Madi, and I. R. Kassem, "Unlocking the potential of DNA microarray for accurate cancer diagnosis with deep learning," in Proc. 7th Int. Conf. Advances in

- Biomedical Engineering (ICABME), 2023, doi: 10.1109/ICABME59496.2023.10293017.
- [5] H. S. Basavegowda and G. Dagnew, "Deep learning approach for microarray cancer data classification," *CAAI Trans. Intell. Technol.*, vol. 5, no. 1, pp. 22–33, 2019, doi: 10.1049/trit.2019.0028.
- [6] S. Gupta, M. K. Gupta, M. Shabaz, and A. Sharma, "Deep learning techniques for cancer classification using microarray gene expression data," *Frontiers in Physiology*, vol. 13, 2022, doi: 10.3389/fphys.2022.952709.
- [7] Y. Zeng, Y. Zhang, Z. Xiao, and H. Sui, "A multi-classification deep neural network for cancer type identification from high-dimension, small-sample and imbalanced gene microarray data," *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-025-89475-2.
- [8] B. C. Feltes, E. B. Chandelier, B. I. Grisci, and M. Dorn, "CuMiDa: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research," *J. Comput. Biol.*, vol. 26, no. 4, pp. 376–386, 2019, doi: 10.1089/cmb.2018.0238.
- [9] M. Ilyas, K. M. Aamir, S. Manzoor, and M. Deriche, "Linear programming based computational technique for leukemia classification using gene expression profile," *PLoS ONE*, vol. 18, no. 10, p. e0292172, 2023, doi: 10.1371/journal.pone.0292172.
- [10] H. Z. Almarzouki, "Deep-learning-based cancer profiles classification using gene expression data profile," *J. Healthcare Eng.*, vol. 2022, pp. 1–13, 2022, doi: 10.1155/2022/4715998.
- [11] S. Sharma, S. Sharma, and A. Athaiya, "Activation functions in neural networks," *Int. J. Eng. Appl. Sci. Technol.*, vol. 4, no. 12, pp. 310–316, 2020, doi: 10.33564/IJEAST.2020.V04I12.054.
- [12] B. Škrlić, S. Džeroski, N. Lavrač, and M. Petković, "Feature importance estimation with self-attention networks," in *Proc. Eur. Conf. Artificial Intelligence (ECAI)*, 2020, pp. 1491–1498. [Online]. Available: <https://dblp.uni-trier.de/db/conf/ecai/ecai2020.html#SkrlicDLP20>
- [13] D. S. R. Polamuri, M. Kumbhkar, and D. A. P. Daniel, *Introduction to Deep Learning*, 1st ed. India: AGPH Books, 2022, ISBN: 978-93-94339-21-7.
- [14] I. Zeger and G. Sisul, "Introduction to deep learning possibilities in communication systems," in *Proc. Int. Symp. ELMAR*, vol. 31, 2021, pp. 21–24, doi: 10.1109/ELMAR52657.2021.9550825.
- [15] A. A. Alnuaim et al., "Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier," *J. Healthcare Eng.*, vol. 2022, pp. 1–12, 2022, doi: 10.1155/2022/6005446.

