JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 6, No. 4. September 2024 DOI: https://doi.org/10.34288/jri.v6i4.350
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

OPTIMIZING DEEP LEARNING WITH DIMENSIONALITY REDUCTION FOR ANALYZING THE CUMIDA BRAIN CANCER GENE EXPRESSION DATASET

Duwi Lufita Marfiana[-1], Fatimah Asmita Rani [-2]

[1,2]Ilmu Komputer /Fakultas Teknologi Informasi
Universitas Nusa Mandiri
[1]Lufita.m98@gmail.com, [2]fasmitarani@gmail.com

Abstract

In the digital era, machine learning and deep learning have become indispensable tools for bioinformatics, particularly in analyzing high-dimensional gene expression data for cancer diagnosis and classification. This study leverages the CuMiDa brain cancer dataset, a curated microarray database with 54,676 genes and 130 samples, to evaluate the effectiveness of deep learning models integrated with dimensionality reduction techniques. Principal Component Analysis (PCA) and Truncated Singular Value Decomposition (TruncatedSVD) were employed to address the challenges of high-dimensional data, reducing noise and computational complexity. Three deep learning models—DNN, MLP, and TabNet—were implemented with various optimizers, including ADAM, RMSprop, and SGD. Results showed that TruncatedSVD outperformed PCA in minimizing loss, especially for MLP with LBFGS optimizers, achieving near-zero loss. TabNet demonstrated the highest classification accuracy (96%) with ADAM and RMSprop. Conversely, SGD exhibited suboptimal performance across models. These findings highlight the critical role of dimensionality reduction and optimizer selection in enhancing the efficiency and accuracy of deep learning models for cancer classification. This research provides a robust framework for improving diagnostic tools in computational oncology.

Keywords: *Dimensionality Reduction; CuMiDa; Brain Cancer; PCA, TruncatedSVD*

Abstrak

Di era digital, machine learning dan deep learning telah menjadi alat yang sangat diperlukan untuk bioinformatika, terutama dalam menganalisis data ekspresi gen berdimensi tinggi untuk diagnosis dan klasifikasi kanker. Penelitian ini memanfaatkan dataset kanker otak CuMiDa, sebuah basis data microarray yang telah dikurasi dengan 54.676 gen dan 130 sampel, untuk mengevaluasi keefektifan model deep learning yang diintegrasikan dengan teknik reduksi dimensi. Principal Component Analysis (PCA) dan Truncated Singular Value Decomposition (TruncatedSVD) digunakan untuk mengatasi tantangan data berdimensi tinggi, mengurangi noise dan kompleksitas komputasi. Tiga model deep learning-DNN, MLP, dan TabNet-diimplementasikan dengan berbagai pengoptimal, termasuk ADAM, RMSprop, dan SGD. Hasil penelitian menunjukkan bahwa TruncatedSVD mengungguli PCA dalam meminimalkan kerugian, terutama untuk MLP dengan pengoptimal LBFGS, mencapai kerugian yang mendekati nol. TabNet menunjukkan akurasi klasifikasi tertinggi (96%) dengan ADAM dan RMSprop. Sebaliknya, SGD menunjukkan kinerja yang kurang optimal di seluruh model. Temuan ini menyoroti peran penting pengurangan dimensi dan pemilihan pengoptimal dalam meningkatkan efisiensi dan akurasi model pembelajaran mendalam untuk klasifikasi kanker. Penelitian ini memberikan kerangka kerja yang kuat untuk meningkatkan alat diagnostik dalam onkologi komputasi

Kata kunci: Dimensionality Reduction; CuMiDa; Kangker Otak; PCA, TruncatedSVD

INTRODUCTION

Along with the advancement of time and the increasing use of technology in daily life, many technologies have emerged that can be applied in various fields of human life, one of which is the health sector. One of the developing digital technologies is machine learning (ML). Machine

learning is a computer system used to perform tasks without direct commands or instructions, but this technology relies on inference patterns or the process of drawing conclusions based on provided evidence or information. Especially in bioinformatics, the shift from traditional methods to ML-based approaches has become very necessary, particularly in gene expression and microarray data analysis, which play a crucial role in the diagnosis and treatment of cancer. (Bagiroz et al., 2020) (Deng & Xu, 2019).

Gene expression analysis allows researchers to uncover patterns and relationships in complex genetic data, which is crucial for understanding cancer development. Microarray technology facilitates the simultaneous measurement of thousands of genes, providing a detailed picture of the genetic profile of cancer cells. However, the high-dimensional nature of microarray datasets—characterized by tens of thousands of genes but relatively few samples—poses challenges such as overfitting, computational inefficiency, and difficulties in extracting biologically relevant patterns (Tabassum et al., 2024) (Deng & Xu, 2019). Traditional machine learning methods, such as Support Vector Machines (SVM) and Random Forests, are widely used but often require extensive manual feature selection and engineering, thereby limiting scalability and adaptability to complex datasets (Chebli et al., 2023).

Deep learning (DL) has emerged as a transformative technology capable of overcoming these limitations. Unlike traditional ML approaches, DL models automatically learn hierarchical representations of raw data, enabling the identification of complex and nonlinear relationships. Convolutional Neural Networks (CNN) and Multilayer Perceptron (MLP) are very effective for analyzing high-dimensional datasets. This demonstrates outstanding performance in cancer classification tasks by dynamically identifying important features and improving predictive accuracy (Younis et al., 2022) (Das et al., 2024). Recent studies show that combining advanced feature selection and dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Mutual Information, can further enhance the performance and interpretability of DL models. (Das et al., 2024) (Basavegowda & Dagnew, 2019).

Brain tumor, one of the most serious and life-threatening diseases. The disease is characterized by the uncontrolled growth of abnormal cells in brain tissue. Early detection and accurate diagnosis will significantly impact the chances of successful treatment. This research proposes a new approach that integrates CNN with the VGG-16 architecture and ensemble models to improve the accuracy of brain tumor detection and classification. The proposed model not only focuses on global optimization but also considers the local characteristics of MRI images, resulting in a more accurate representation. By using a dataset consisting of 253 brain MRI images, including 155 tumor cases, this research aims to develop a system that can assist medical professionals in making faster and more accurate diagnostic decisions (Younis et al., 2022).

Curated Microarray Database (CuMiDa) is an important resource for comparing machine learning and deep learning models in cancer analysis. This extensively curated dataset includes high-quality microarray data that has undergone thorough preprocessing, ensuring noise removal and data standardization. The brain cancer subset from CuMiDa, which consists of 54,676 genes and 130 samples across five classes, represents an ideal testbed for evaluating DL models in the context of high-dimensional data analysis (Ilyas et al., 2023) (Feltes et al., 2019). Studies using CuMiDa have highlighted its utility in developing robust models for gene expression analysis and cancer classification, providing a foundation for advancing computational oncology (Feltes et al., 2019).

The study by Ilyas (Ilyas et al., 2023) used a leukemia gene expression dataset from the CuMiDa. The research explains how the use of microarrays is an ideal approach for studying cancer. Based on the context of deep learning development and research, CuMiDa plays an important role as a high-quality data source for analyzing gene expression in cancer classification. The increasing complexity of data and the need for large and diverse datasets make CuMiDa a source of extensively curated cancer microarray datasets, allowing researchers to focus on developing deep learning models without worrying about the quality and consistency of the data. The datasets available in CuMiDa have undergone thorough pre-processing, including normalization and noise removal, which are used to enhance the accuracy of deep learning models. With a standardized dataset, researchers can more easily apply deep learning techniques, such as CNN. The technique is used to identify complex patterns in gene expression data. Based on that statement, it can be concluded that CuMiDa not only provides access to high-quality data but also contributes to the

development and validation of deep learning models in cancer analysis (Feltes et al., 2019).

In addition to impacting bioinformatics, DL has gained traction in modern communication systems, where DL addresses significant challenges such as varying channel conditions and the complexity of real-world scenarios. The phenomenon utilizes architecture inspired by the human brain. DL has introduced an innovative approach to optimize this system. For example, autoencoders, a type of unsupervised neural network, use a communication process by integrating transmitter and receiver functions, thereby eliminating traditional block structures and optimizing metrics such as block error rates. This advancement reduces reliance on handcrafted models while simultaneously increasing automation and efficiency. However, such challenges must be adjusted to the needs of extensive datasets and effective computational resources to ensure guaranteed quality generalization across various applications (Zeger & Sisul, 2021) (Polamuri et al., 2022).

The increasing need for accurate and efficient diagnostic tools in cancer research also drives the exploration of hybrid methods and new optimization techniques. An example of this phenomenon is how feature selection based on genetic algorithms shows a significant improvement in classification results for high-dimensional microarray data (Ali & Saeed, 2023) (Nagra et al., 2024). Additionally, advancements in activation functions and optimization strategies have played a crucial role in improving the performance of neural networks, enabling them to tackle issues such as vanishing gradients and overfitting (Lederer, 2021) (Sharma et al., 2020).

This research focuses on managing the risks posed by high-dimensional data and improving classification accuracy using DL models tailored to the CuMiDa dataset. This research integrates dimensionality reduction methods and feature selection techniques with DL frameworks such as CNN and MLP. The aim of this research is to reduce the curse of dimensionality while maintaining the biological relevance of the features. Additionally, ensemble approaches and optimization techniques are combined to refine the model, allowing for better generalization and accurate predictions (Das et al., 2024) (Basavegowda & Dagnew, 2019) (Gupta et al., 2022).

RESEARCH METHODS

The stages carried out in this research are illustrated in Figure 1, the research flowchart. The diagram shows nine stages using the literature review and research approach. Data collection in this study includes data preprocessing, data splitting, feature reduction, data augmentation, model implementation, model training, and cross-validation. The explanation of each stage is as follows:

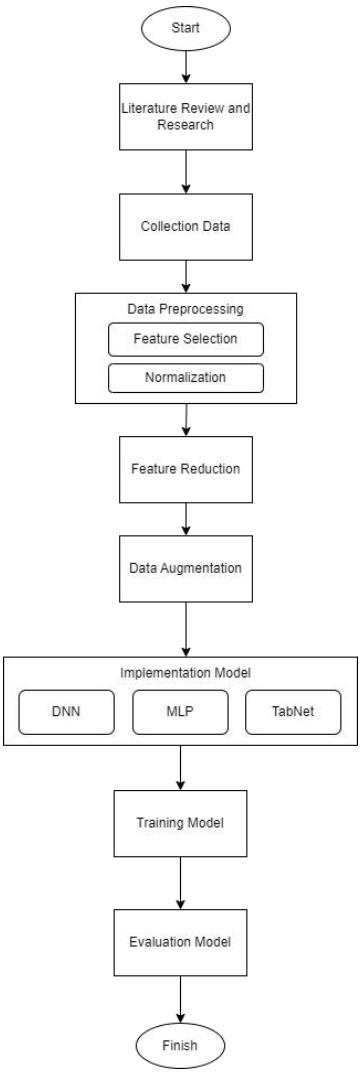

Figure 1. Flowchart Research

1. Literature Review

This research involves a comprehensive review of the literature and methodologies related to deep learning, dimensionality reduction, and the CuMiDa brain cancer gene expression dataset (Feltes et al., 2019). This step establishes a solid theoretical framework to guide the research.

2. Collecting Data

The dataset used in this research comes from CuMiDa, a microarray database. Microarray technology is crucial for analyzing gene expression profiles, as it enables the identification of patterns associated with various types of cancer (Tabares-Soto et al., 2020) (Basavegowda & Dagnew, 2019).

3. Data Preprocessing

At this stage, the focus is on refining the CuMiDa brain cancer gene expression dataset to ensure that the selected features are informative so the data is ready for effective model training using feature selection methods such as Variance Threshold and ANOVA, as well as normalization.

The Variance Threshold method evaluates the variance of each feature to determine whether that feature should be retained. Variance, denoted as σ^2, measures the spread of data for a particular feature. It is calculated as the average squared difference between each data point x_i and the mean μ of the feature, divided by the total number of samples n. The equation for variance is:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2 \quad (1)$$

The ANOVA method identifies features that show statistically significant differences across different classes. It can be ensured that only the most relevant features can be retained, which is very important to avoid overfitting and thus accelerate model training (Nasiri & Alavi, 2022). This method uses the F-statistic, which is the ratio of the variance between group (S_B^2) and the variance within group (S_W^2). The F-statistic is calculated as:

$$F = \frac{S_B^2}{S_W^2} \quad (2)$$

Normalization is a key preprocessing step to ensure all dataset features have a uniform scale. By using techniques such as min-max scaling, the values are transformed into a standard range, typically [0, 1], by identifying the minimum value (x_{min}) and the maximum value (x_{max}) of each feature and applying the formula:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

This approach is very important for high-dimensional datasets like CuMiDa, where the features often show significant scale variations. This approach also enhances the convergence and stability of the model, especially for deep learning algorithms that are sensitive to input scale (Singh & Singh, 2019).

4. Splitting Data

The dataset is divided into two subsets, 80% for training the model and 20% for testing its performance. This division ensures that the model is trained effectively while retaining a portion of the data for unbiased evaluation. The training set is used to develop and optimize the model by learning patterns and relationships in the data.

5. Reduction Dimension Feature

After preprocessing and data separation, dimensionality reduction is applied to reduce the dimensions of the CuMiDa brain cancer gene expression dataset. Dimensionality reduction is crucial for handling high-dimensional datasets like CuMiDa, which may contain redundant or irrelevant features. PCA transforms the original features into a smaller set of uncorrelated variables, called principal components, while retaining as much variability as possible from the original data.

This projection is done by multiplying the data with the eigenvector matrix. The new data representation is obtained using:

$$x_{new} = U^T x \quad (9)$$

U is a matrix that contains the upper part of the eigenvector, x is the standardized data vector, and x_{new} is the resulting value, a lower-dimensional representation of the data. This step reduces the dataset's dimensions while preserving the most important information from the original features, as represented by the largest eigenvalues.

The Truncated Singular Value Decomposition (Truncated SVD) method is a dimensionality reduction technique commonly used to mitigate the challenges associated with high-dimensional data in various applications, including indoor positioning. This approach is based on the concept of Singular Value Decomposition (SVD), which decomposes the data matrix into three separate matrices representing the singular values and the corresponding left and right singular vectors (Thi et al., 2023).

While PCA works by identifying directions (principal components) that maximize variance, TruncatedSVD decomposes the data matrix into singular values and associated vectors. Then it only retains the top singular values, effectively reducing the number of dimensions while preserving the most significant features of the data.

$$A = U\Sigma V^T \quad (10)$$

A is the original data matrix, U is an $m \times r$ orthogonal matrix, Σ is an $r \times r$ orthogonal matrix with singular values on the diagonal, and V^T is an $r \times n$ orthogonal matrix. In TruncatedSVD, the number of singular values r is reduced by retaining only the top k values. The rank k is chosen based on the proportion of variance we want to retain in the data.

6. Data Augmentation

Data augmentation is a technique to artificially enhance the training set by creating modified copies of the dataset using existing data. One of the augmentation methods is SMOTE. The Synthetic Minority Over-sampling Technique (SMOTE) is used to classify imbalanced datasets. This technique synthesizes new samples from the minority class to balance the dataset by resampling from the minority class examples (Brandt & Lanzén, 2020).

7. Model Implementation

In this study, after splitting the data, applying dimensionality reduction, and performing SMOTE, the deep learning stage was implemented using three models: DNN, TabNet, and MLP, each evaluated with several optimizers to compare their performance under the same conditions. The DNN architecture typically includes an input layer, several hidden layers, and an output layer. In this study, the DNN architecture includes a 64-unit Dense layer with ReLU activation and L2 regularization (alpha = 0.01) to prevent overfitting, followed by a Dropout layer with a rate of 0.2 to reduce overfitting. The second dense layer with 32 units, ReLU activation, and L2 regularization captures additional non-linear relationships. The output layer consists of 5 units with softmax activation, which is suitable for multi-class classification tasks. DNNs are known for their ability to learn complex patterns and generalize well to unseen data when properly tuned.

The TabNet model, implemented using PyTorch, is a deep learning model designed to efficiently handle tabular data. This model is based on decision trees and an attention mechanism, allowing it to focus on relevant features while maintaining interpretability. TabNet is configured with specific hyperparameters: n_d (decision step output dimension) is set to 32, n_a (attention output dimension) is also set to 32, and n_steps (number of decision steps) is set to 5. These parameters determine the model's capacity to pay attention to relevant features at each decision step. n_d allows the model to learn more complex patterns, while n_a defines the attention mechanism that helps focus on important features for easier interpretation and efficiency. The n_steps parameter controls the number of decision steps taken by the model during training, allowing the model to explore more feature combinations. Gamma, set to 1.5, controls the entropy of the decision tree, and lambda_sparse, set to 1e-3, is used for regularization to encourage sparsity in the feature selection process. TabNet was trained using the Adam, RMSprop, and SGD optimizers, and this model used a batch size of 8 and a virtual batch size of 4 to optimize memory efficiency. This model was trained with early stopping to prevent overfitting, ensuring effective learning and generalization.

MLP (Multi-Layer Perceptron) is a class of artificial neural networks consisting of several layers of neurons, where each neuron in one layer is connected to every neuron in the next layer. The MLP used in this study is implemented using MLPClassifier from scikit-learn, which consists of three hidden layers with a decreasing number of neurons: 128, 64, and 32, each using ReLU activation to effectively capture non-linear relationships in the data. L2 regularization with an alpha value of 0.01 is applied to prevent overfitting by penalizing large weights. The MLP model was tested with three different optimizers: Adam, LBFGS, and SGD. Adam is a powerful optimizer that adapts the learning rate based on the gradient, making it suitable for complex data. LBFGS, a quasi-Newton method, ensures rapid convergence and effective use of second-order information, making it very helpful for more complex optimization problems. SGD offers a baseline for comparison by updating weights with a constant learning rate, although it may be less adaptive compared to other optimizers. The MLP model also uses early stopping with validation_fraction = 0.2

to reserve a small portion of the training data for validation and prevent overfitting during training.

These three models—DNN, TabNet, and MLP—are all neural network-based architectures that learn from data through different mechanisms, with DNN utilizing several dense layers, TabNet focusing on attention steps and decisions, and MLP using fully connected layers with backpropagation and optimization. Each model is carefully configured and evaluated to optimize performance for multi-class classification tasks.

8. Training Model

In the model training phase, the DNN model is trained for a maximum of 100 epochs, with early stopping applied to monitor validation loss. Training is stopped if no improvement is observed for 10 epochs to avoid overfitting, and the best weights are restored. The batch size is set to 32, and class weights are used to address class imbalance in the training data.

The MLP model was trained with a maximum of 200 iterations, and early stopping was activated to monitor performance on the validation subset, which consists of 20% of the training data. The TabNet model was trained for a maximum of 100 epochs, with early stopping applied after 100 epochs without improvement to prevent overfitting. The batch size is set to 8, and the virtual batch size is set to 4 for optimized memory usage during training.

9. Model Evaluation

At the model evaluation stage, the model's performance is assessed using several metrics to determine how effective the model is in making predictions on unseen data. During training, the maximum validation accuracy (max_accuracy) and minimum validation loss (min_loss) are tracked for each fold. These values are then stored and printed for each fold, providing a measure of how well the model performed during training. Specifically, max_accuracy represents the highest validation accuracy achieved during training, while min_loss indicates the lowest loss observed. These metrics are useful for understanding the model's ability to learn from data and avoid overfitting.

To gain deeper insights into the model's classification performance, a confusion matrix and classification report were created. The confusion matrix provides detailed information about true positives, false positives, true negatives, and false negatives, which helps evaluate how effectively the model distinguishes between classes. The classification report offers additional metrics such as precision, recall, and F1-score for each class, providing a more comprehensive view of the model's performance. The model's predictions are compared with the actual test labels, and the confusion matrix and classification report are printed for each fold.

Finally, the results of the cross-validation process are summarized by calculating the average accuracy, standard deviation of accuracy, average maximum accuracy, and average minimum loss across all folds. These metrics offer a broader view of the model's performance, providing an understanding of its consistency and generalization capability across different data splits. The average accuracy and standard deviation of accuracy are crucial for evaluating the overall effectiveness of the model, while the maximum average accuracy and minimum loss during training provide insights into how well the model learns from the training data. This comprehensive evaluation process ensures that the model's performance is thoroughly assessed, highlighting strengths and areas with potential for improvement.

RESULTS AND DISCUSSION

In this study, three deep learning models, namely DNN, MLP, and TabNet, were tested on the CuMiDa dataset. This dataset has high-dimensional characteristics and complex patterns, requiring efficient data processing methods. This research focuses on evaluating model performance based on the combination of several optimizers (ADAM, RMSPROP, LBFGS, and SGD) and the application of dimensionality reduction techniques, namely Principal Component Analysis (PCA) and Truncated Singular Value Decomposition (TruncatedSVD). The evaluation was conducted by comparing the accuracy and loss of the model with dimensionality reduction.

The results obtained show that there is a variation in the impact of dimensionality reduction on model performance, but this depends on the type of model and optimization used. Some combinations of models and dimensionality reduction techniques show an increase in learning efficiency, both in terms of accuracy and loss reduction. While some other combinations do not have a significant impact or even reduce performance. The following discussion details the

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v6i4.350
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

performance of each model based on the test results, as well as the comparison between models in the context of the CuMiDa dataset.

Table 1. Accuracy and Loss Metrics with PCA Reduction

No	Model	Optimizer	Accuracy	Loss
1	DNN	ADAM	0.95	0.42
		RMSprop	0.95	0.26
		SGD	0.94	0.34
2	MLP	ADAM	0.92	0.4
		LBFGS	0.95	0.04
		SGD	0.85	2.37
3	TabNet	ADAM	0.96	0.23
		RMSprop	0.96	0.24
		SGD	0.3	4.43

In the PCA-based testing in Table 1, the DNN model with ADAM, RMSprop, and SGD optimizers showed stable accuracy between 0.94 and 0.95. The application of PCA seems to contribute to the reduction of loss, especially in the DNN model with RMSprop, which resulted in the lowest loss of 0.26. On the other hand, in the MLP, although the accuracy did not experience a significant increase (there was a slight decrease in accuracy with SGD at 0.85), the use of LBFGS resulted in the lowest loss of 0.04. The SGD optimizer on MLP struggled with the application of PCA, resulting in a very high loss (2.37). The TabNet model shows the best performance with an accuracy of 0.96 and a loss of 0.23 using the ADAM and RMSprop optimizers. However, the use of the SGD optimizer showed poor performance with an accuracy of only 0.30 and a very high loss (4.43). Overall, the application of PCA provides a significant improvement for the TabNet and DNN models, but the results for the MLP with the SGD optimizer are less optimal.

Table 2. Accuracy and Loss Metrics with TruncatedSVD Reduction

No	Model	Optimizer	Accuracy	Loss
1	DNN	ADAM	0.94	0.37
		RMSprop	0.95	0.37
		SGD	0.95	0.34
2	MLP	ADAM	0.92	0.08
		LBFGS	0.92	0.01
		SGD	0.90	0.18
3	TabNet	ADAM	0.96	0.27
		RMSprop	0.96	0.31
		SGD	0.42	3.5

In Table 2, using the TruncatedSVD dimensionality reduction technique, the results show the same trend as the PCA application, but with some differences. The DNN model using the ADAM and RMSprop optimizers still maintains an accuracy of 0.95 with a slightly lower loss (0.37 for RMSprop) compared to PCA. However, in the MLP, the application of TruncatedSVD yielded better results with the lowest loss of 0.08 using the ADAM optimizer, significantly lower than PCA (0.4). LBFGS on MLP also shows improvement with a loss of 0.01, which is the best result among all models in Table 2. The SGD optimizer on MLP resulted in a loss of 0.18 with an accuracy of 0.92. In TabNet, the application of TruncatedSVD resulted in slightly higher losses (0.27 for ADAM and 0.24 for RMSprop), but its accuracy remained optimal at 0.96. The SGD optimizer once again showed poor results, with an accuracy of only 0.42 and a high loss of 3.5. Overall, TruncatedSVD provides a significant improvement in terms of loss, especially for the MLP model, with the LBFGS optimizer resulting in almost zero loss.

Looking at the comparison between Table 1 and Table 2, it is evident that the use of TruncatedSVD (Table 2) yields more stable and better results in reducing loss compared to PCA (Table 1), especially in the MLP model. MLP with LBFGS on TruncatedSVD shows a very low loss (0.01), indicating that this dimensionality reduction technique is very effective in improving the model's efficiency. On the other hand, TabNet shows high accuracy on both tables (0.96) with slightly lower loss using PCA than TruncatedSVD. Meanwhile, in the DNN, both PCA and TruncatedSVD yielded almost identical results in terms of accuracy and loss, indicating that this model is not significantly affected by the application of dimensionality reduction, figure 2 and 3 shows PCA result and Truncated SVD result.

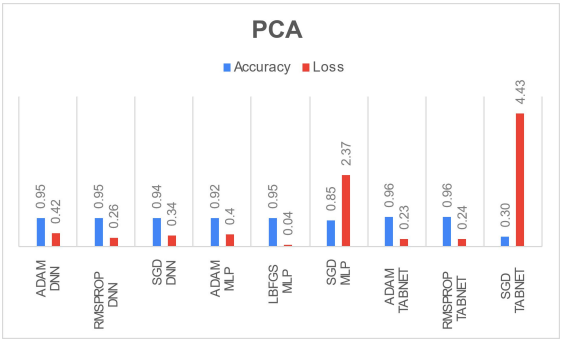

Figure 2. PCA Result

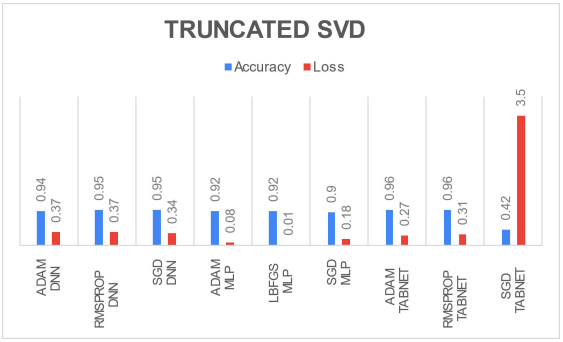

Figure 3. Truncated SVD Result

The SGD optimizer, which performed poorly on both tables, recorded very low accuracy and very high loss with PCA and TruncatedSVD. This indicates that the use of SGD on the CuMiDa dataset may not be optimal, especially with the application of dimensionality reduction techniques.

CONCLUSIONS AND SUGGESTIONS

Conclusions

The application of dimensionality reduction techniques (PCA and TruncatedSVD) improves model performance, with TruncatedSVD tending to provide more stable results than PCA, especially in reducing loss and increasing accuracy. The DNN and TabNet models show better performance than the MLP, with TabNet achieving the highest accuracy. The ADAM and RMSprop optimizers are more effective in optimizing the model, while SGD tends to reduce performance, especially on TabNet. Overall, choosing the right reduction techniques and optimizers has a significant impact on model performance.

Suggestions

Future research is expected to explore the use of more varied techniques and a greater number of sampling methods to improve model performance on more complex datasets. Additionally, testing the AugOS-CNN model on various datasets with higher or lower levels of imbalance could provide further insights into the effectiveness of this method in a broader context. Future studies could also optimize this model by applying regularization techniques or more advanced network architectures to recover potential fit and improve performance on unused data.

REFERENCES

Bagiroz, B., Doruk, E., & Yildiz, O. (2020). Machine learning in Bioinformatics: gene expression and microarray studies. *2020 Medical Technologies Congress (TIPTEKNO)*. https://doi.org/10.1109/tiptekno50054.2020.9299285

Tabassum, N., Kamal, M. a. S., Akhand, M. a. H., & Yamada, K. (2024). Cancer Classification from Gene Expression Using Ensemble Learning with an Influential Feature Selection Technique. *BioMedInformatics*, *4*(2), 1275–1288. https://doi.org/10.3390/biomedinformatics4020070

Deng, X., & Xu, Y. (2019). Cancer Classification Using Microarray Data By DPCAForest. *2019 IEEE 31st International Conference on Tools With Artificial Intelligence (ICTAI)*, 1081–1087. https://doi.org/10.1109/ictai.2019.00151

Chebli, H., Mashhadieh, Z., Ali, M. A., Madi, M. K., & Kassem, I. R. (2023). Unlocking the potential of DNA microarray for accurate cancer diagnosis with deep learning. *2023 Seventh International Conference on Advances in Biomedical Engineering (ICABME)*, 251–256. https://doi.org/10.1109/icabme59496.2023.10293017

Younis, A., Qiang, L., Nyatega, C. O., Adamu, M. J., & Kawuwa, H. B. (2022). Brain tumor analysis using deep learning and VGG-16 ensembling learning approaches. *Applied Sciences*, *12*(14), 7282. https://doi.org/10.3390/app12147282

Tabares-Soto, R., Orozco-Arias, S., Romero-Cano, V., Bucheli, V. S., Rodríguez-Sotelo, J. L., & Jiménez-Varón, C. F. (2020). A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Computer Science*, *6*, e270. https://doi.org/10.7717/peerj-cs.270

Das, A., Neelima, N., Deepa, K., & Özer, T. (2024). Gene selection based cancer classification with adaptive optimization using deep learning architecture. *IEEE Access*, *12*, 62234–62255. https://doi.org/10.1109/access.2024.3392633

Basavegowda, H. S., & Dagnew, G. (2019). Deep learning approach for microarray cancer data classification. *CAAI Transactions on Intelligence Technology*, *5*(1), 22–33. https://doi.org/10.1049/trit.2019.0028

Ilyas, M., Aamir, K. M., Manzoor, S., & Deriche, M. (2023). Linear programming based computational technique for leukemia classification using gene expression profile. *PLoS ONE*, *18*(10), e0292172. https://doi.org/10.1371/journal.pone.0292172

Feltes, B. C., Chandelier, E. B., Grisci, B. I., & Dorn, M. (2019). CUMIDA: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, *26*(4), 376–386. https://doi.org/10.1089/cmb.2018.0238

Zeger, I., & Sisul, G. (2021). Introduction to deep learning possibilities in communication systems. *International Symposium ELMAR*, 21–24. https://doi.org/10.1109/elmar52657.2021.9550825

Polamuri, S. R., Kumbhkar, M., & Daniel, D. A. P. (2022). *Introduction to Deep Learning* (1st Edition). AGPH Books (Academic Guru Publishing House). ISBN: 978-93-94339-21-7.

Gupta, S., Gupta, M. K., Shabaz, M., & Sharma, A. (2022). Deep learning techniques for cancer classification using microarray gene expression data. *Frontiers in Physiology*, *13*. https://doi.org/10.3389/fphys.2022.952709

Ali, W., & Saeed, F. (2023). Hybrid filter and Genetic Algorithm-Based feature selection for improving cancer classification in High-Dimensional Microarray data. *Processes*, *11*(2), 562. https://doi.org/10.3390/pr11020562

Nagra, A. A., Khan, A. H., Abubakar, M., Faheem, M., Rasool, A., Masood, K., & Hussain, M. (2024). A gene selection algorithm for microarray cancer classification using an improved particle swarm optimization. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-68744-6

Lederer, J. (2021). Activation Functions in Artificial Neural Networks: A Systematic Overview. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2101.09957

Sharma, S., Sharma, S., & Athaiya, A. (2020). ACTIVATION FUNCTIONS IN NEURAL NETWORKS. *International Journal of Engineering Applied Sciences and Technology*, *04*(12), 310–316. https://doi.org/10.33564/ijeast.2020.v04i12.054

Nasiri, H., & Alavi, S. A. (2022). A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-Ray Images. *Computational Intelligence and Neuroscience*, *2022*, 1–11. https://doi.org/10.1155/2022/4694567

Singh, D., & Singh, B. (2019). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Brandt, J., & Lanzén, E. (2020). *A comparative review of SMOTE and ADASYN in imbalanced data classification*. https://www.diva-portal.org/smash/get/diva2:1519153/FULLTEXT01.pdf

Thi, H. D., Manh, K. H., Anh, V. T., Quynh, T. P. T., & Viet, T. N. (2023). Dimensionality Reduction with Truncated Singular Value Decomposition and K-Nearest Neighbors Regression for Indoor Localization. *International Journal of Advanced Computer Science and Applications*, *14*(10). https://doi.org/10.14569/ijacsa.2023.0141034