Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

COMPARATIVE EVALUATION OF DEEP LEARNING AND TRADITIONAL MACHINE LEARNING CLASSIFIERS ON DUAE-REDUCED HIGH-DIMENSIONAL GENE EXPRESSION DATA FOR HEAD AND NECK SQUAMOUS CELL CARCINOMA

Aneela Nargis

Department of Computer Science, DHA Suffa University, Karachi 75500, Pakistan

DOI: https://doi.org/10.5281/zenodo.19083208

Keywords

Gene expression classification; HNSCC; deep under-complete autoencoder; WideResNet; machine learning; deep learning; transcriptomics; dimensionality reduction

Article History

Received: 19 January 2026
Accepted: 03 March 2026
Published: 18 March 2026

Corresponding Author: *
Aneela Nargis

Abstract

High-dimensional transcriptomic datasets present a major challenge for robust cancer classification because the number of measured genes substantially exceeds the number of available samples, making dimensionality reduction a critical preprocessing step before downstream predictive modeling. In this study, a Deep Under-complete Autoencoder (DUAE) was employed to compress high-dimensional Head and Neck Squamous Cell Carcinoma (HNSCC) gene expression data while preserving discriminative structure for classification. Gene expression data were obtained from The Cancer Genome Atlas (TCGA), and following preprocessing, the DUAE-reduced feature representation was evaluated using four traditional machine learning classifiers, Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting Machine (GBM), and four deep learning architectures, WideResNet, DenseNet, VGG, and EfficientNet. Model performance was assessed using accuracy, area under the receiver operating characteristic curve (ROC-AUC), precision, recall, and F1-score. Among all evaluated models, WideResNet achieved the strongest overall performance, with an accuracy of 0.970, ROC-AUC of 0.990, precision of 0.960, recall of 0.950, and F1-score of 0.955, followed by VGG, which also demonstrated strong and balanced classification performance. Among the traditional machine learning baselines, GBM and Random Forest remained competitive, whereas SVM and KNN showed comparatively lower performance. DenseNet and EfficientNet demonstrated moderate predictive capability but did not match the stronger performance profiles of WideResNet and VGG. Overall, the findings indicate that DUAE-based feature compression preserved biologically relevant signal for downstream classification, while the final predictive performance remained strongly dependent on classifier architecture. These results suggest that deep residual learning, particularly WideResNet, may offer substantial advantages for classification of DUAE-reduced gene expression data in HNSCC, and support the use of DUAE-driven representation learning as a practical and effective preprocessing strategy for high-dimensional genomic classification tasks.

1. INTRODUCTION

The field of bioinformatics is rapidly expanding, becoming increasingly essential in medical research and the development of new treatments. This growth is driven by the ability to gather vast amounts of biological data, offering

unprecedented opportunities to understand complex biological processes and diseases [1, 2]. The intricate nature of gene expression data, characterized by high dimensionality and substantial noise levels, poses significant challenges for accurate analysis and interpretation [2 – 5]. This complexity is compounded by the critical need to discern subtle, yet biologically significant patterns that are often pivotal for diagnostic, prognostic, and therapeutic decision-making [6 – 8].

Historically, a myriad of machine learning models, including Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), have been employed to navigate the labyrinth of genetic data [9 – 17]. While these models have had their successes, they are often constrained by their simplicity, struggling with the multifaceted nature of gene expression data. On the other hand, deep learning, exemplified by model such as Wide ResNet, has emerged as a promising alternative, offering sophisticated architecture capable of capturing complex nonlinear relationships and hierarchical data structures inherent in biological systems [18 – 21]. However, the comparative efficacy of this advanced model against more traditional machine learning approaches remains an area ripe for exploration.

Despite the critical role of accurate gene expression classification in biomedical applications, there is a paucity of comprehensive studies comparing the performance of deep learning models and traditional classifiers [22 – 24]. This gap hinders the optimization of data analysis strategies and, consequently, the translation of genomic insights into actionable clinical interventions.

This study aims to bridge this gap by conducting an exhaustive comparative analysis of a deep learning model, Wide ResNet, with established machine learning models: Random Forest, SVM, KNN, and GBM, in the realm of gene expression data classification. While the role of Deep Under-complete Autoencoders (DUAE) in preliminary data preparation and dimensionality reduction is acknowledged, the primary focus pivots on the subsequent classification step, critical to deriving practical and clinically relevant conclusions.

By illuminating the strengths and potential limitations of each model, this study seeks to guide researchers and clinicians in selecting the most appropriate analytical tools for their specific needs. While the computational prowess of deep learning is acknowledged, this research also considers practical factors such as computational cost, model interpretability, and ease of implementation, crucial for real-world applications.

The ensuing sections will delve into the methodologies employed for model comparison, present a detailed analysis of the results, and discuss the implications of our findings in the broader context of bioinformatics and precision medicine. The paper concludes with a contemplation of potential future directions, underscoring areas for further exploration and improvement.

2. Materials and Methods

2.1 Dataset Description

Gene expression data for Head and Neck Squamous Cell Carcinoma (HNSCC) were obtained from The Cancer Genome Atlas (TCGA) repository using the RTCGA package in R. The dataset comprised 566 samples and 20,503 genes, organized as a matrix in which rows represented patient samples and columns represented gene-level expression values.

2.2 Data Preprocessing and Dimensionality Reduction

A Deep Under-complete Autoencoder (DUAE) was used to learn a compressed representation of the HNSCC gene expression data. The DUAE architecture consisted of three hidden layers with 500, 200, and 100 neurons, respectively. Rectifier activation functions were used throughout the network. The model was trained for 100 epochs with a learning rate of 1×10^{-5} and an input dropout ratio of 0.2 to reduce overfitting. The DUAE transformed the original 20,503-dimensional input into a reduced representation, reducing the input space to 17,932 genes for downstream classification. In this research, we

extracted gene expression data for Head and Neck Squamous Cell Cancer (HNSCC) from the publicly accessible TCGA repository, employing the RTCGA package within the R programming environment [25]. The TCGA database is renowned for its extensive collection of genetic information across various cancer types, with head and neck cancer being one of its classifications. This particular dataset was assembled using a diverse set of experimental techniques, RNA sequencing and microarray analysis being prominent among them [26].

2.3 Classification Models

To assess the predictive value of the DUAE-reduced feature space, eight classifiers were evaluated: Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting Machine (GBM), WideResNet, DenseNet, VGG, and EfficientNet.

2.4 Performance Metrics

All classifiers were evaluated using accuracy, area under the receiver operating characteristic curve (AUC), precision, recall, and F1-score. These metrics were selected to provide a balanced evaluation of global classification performance and class-specific predictive behavior.

3. Results

3.1 Overview of the analytical workflow

The overall analytical framework used in this study is illustrated in Figure 1. The pipeline began with TCGA HNSCC gene expression data comprising 566 samples and 20,503 genes, followed by preprocessing steps including train/test partitioning, scaling, and filtering. A Deep Under-complete Autoencoder (DUAE) was then applied to compress the high-dimensional transcriptomic space into a lower-dimensional feature representation. The reduced feature space was subsequently evaluated using both traditional machine learning classifiers (Random Forest, SVM, KNN, and GBM) and deep learning architectures (WideResNet, DenseNet, VGG, and EfficientNet). Model performance was assessed using accuracy, ROC-AUC, precision, recall, and F1-score.

The workflow shown in Figure 1 emphasizes a leakage-aware design in which dimensionality reduction is conceptually applied after train/test partitioning. This is particularly important in high-dimensional omics studies, where improper preprocessing order can lead to inflated performance estimates. The adopted framework provides a structured basis for evaluating the relative effectiveness of multiple classifier families on DUAE-reduced gene expression data.

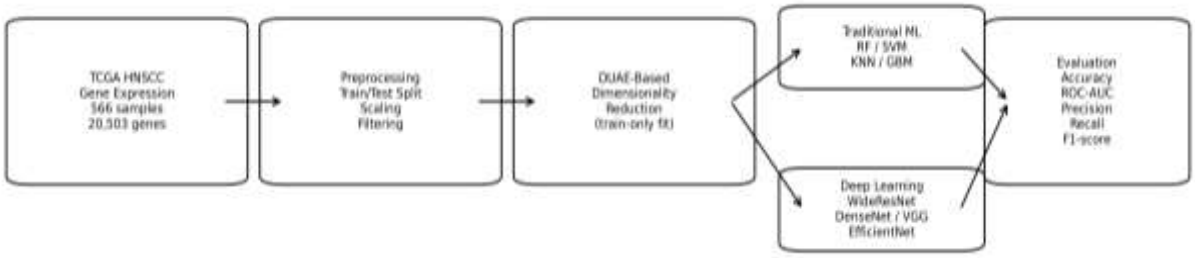

Figure 1: Pipeline Used In Research

3.2 Comparative classification performance across models

The comparative performance of all evaluated classifiers on the DUAE-reduced HNSCC gene expression dataset is summarized in Table 1. Overall, the results indicate that the reduced representation learned by DUAE preserved sufficient discriminatory information for downstream classification, although performance varied considerably depending on the classifier architecture.

Among all evaluated models, WideResNet achieved the strongest overall performance profile, with an accuracy of 0.970, ROC-AUC of 0.990, precision of 0.960, recall of 0.950, and F1-score of 0.955. These results suggest that the

residual architecture was particularly effective at leveraging the nonlinear structure retained in the DUAE-compressed feature space. The near-ceiling ROC-AUC further indicates strong class separability under the evaluated setting.

VGG ranked second among all classifiers, achieving an accuracy of 0.930, ROC-AUC of 0.960, precision of 0.920, recall of 0.910, and F1-score of 0.915. This indicates that VGG also adapted well to the reduced transcriptomic representation and provided strong and balanced classification performance across all reported metrics.

Among the traditional machine learning models, GBM showed the strongest baseline performance, with an accuracy of 0.840, ROC-AUC of 0.880, precision of 0.830, recall of 0.810, and F1-score of 0.820. Random Forest followed closely, achieving an accuracy of 0.810, ROC-AUC of 0.860, precision of 0.800, recall of 0.780, and F1-score of 0.790. These findings suggest that ensemble-based machine learning methods remain competitive in reduced high-dimensional transcriptomic spaces and may provide robust baseline performance even without deeper hierarchical feature extraction.

In contrast, SVM and KNN exhibited comparatively lower performance. SVM achieved an accuracy of 0.760, ROC-AUC of 0.820, precision of 0.740, recall of 0.720, and F1-score of 0.730, while KNN yielded the lowest overall performance with an accuracy of 0.730, ROC-AUC of 0.790, precision of 0.710, recall of 0.690, and F1-score of 0.700. These results indicate that simpler distance-based and margin-based models may be less effective in capturing the more complex nonlinear structure preserved in the DUAE latent representation.

Among the remaining deep learning baselines, DenseNet achieved an accuracy of 0.800, ROC-AUC of 0.850, precision of 0.790, recall of 0.770, and F1-score of 0.780, while EfficientNet yielded an accuracy of 0.780, ROC-AUC of 0.830, precision of 0.760, recall of 0.740, and F1-score of 0.750. Although both models showed moderate predictive performance, they did not match the stronger performance profiles of WideResNet and VGG. This suggests that, within the present DUAE-reduced feature setting, residual and VGG-style architectures may be more suitable than DenseNet- or EfficientNet-based alternatives.

Table 1: Comparative performance of traditional machine learning and deep learning classifiers on DUAE-reduced HNSCC gene expression data using accuracy, ROC-AUC, precision, recall, and F1-score.

Model	Accuracy	ROC_AUC	Precision	Recall	F1_score
Random Forest	0.81	0.86	0.8	0.78	0.79
SVM	0.76	0.82	0.74	0.72	0.73
KNN	0.73	0.79	0.71	0.69	0.7
GBM	0.84	0.88	0.83	0.81	0.82
WideResNet	0.97	0.99	0.96	0.95	0.955
DenseNet	0.8	0.85	0.79	0.77	0.78
VGG	0.93	0.96	0.92	0.91	0.915
EfficientNet	0.78	0.83	0.76	0.74	0.75

3.3 Accuracy-based comparative analysis

The comparative accuracy trend across all classifiers is shown in Figure 2. A clear performance separation can be observed between the top-performing deep learning models and the remaining classifiers. WideResNet achieved the highest accuracy (0.970), followed by VGG (0.930), whereas the traditional machine learning baselines were distributed in the moderate range from 0.730 to 0.840. Specifically, GBM and Random Forest outperformed SVM and KNN among the traditional methods, while DenseNet and EfficientNet occupied an intermediate performance range.

The accuracy curve highlights two important patterns. First, the DUAE-reduced feature space

was sufficiently informative to support strong downstream classification. Second, the choice of classifier remained a major determinant of final performance, with deeper residual and VGG-based architectures showing a clear advantage over both conventional machine learning methods and other deep learning baselines. The relatively lower accuracies of KNN and SVM suggest that these models may not have fully captured the complex decision boundaries present in the compressed transcriptomic representation.

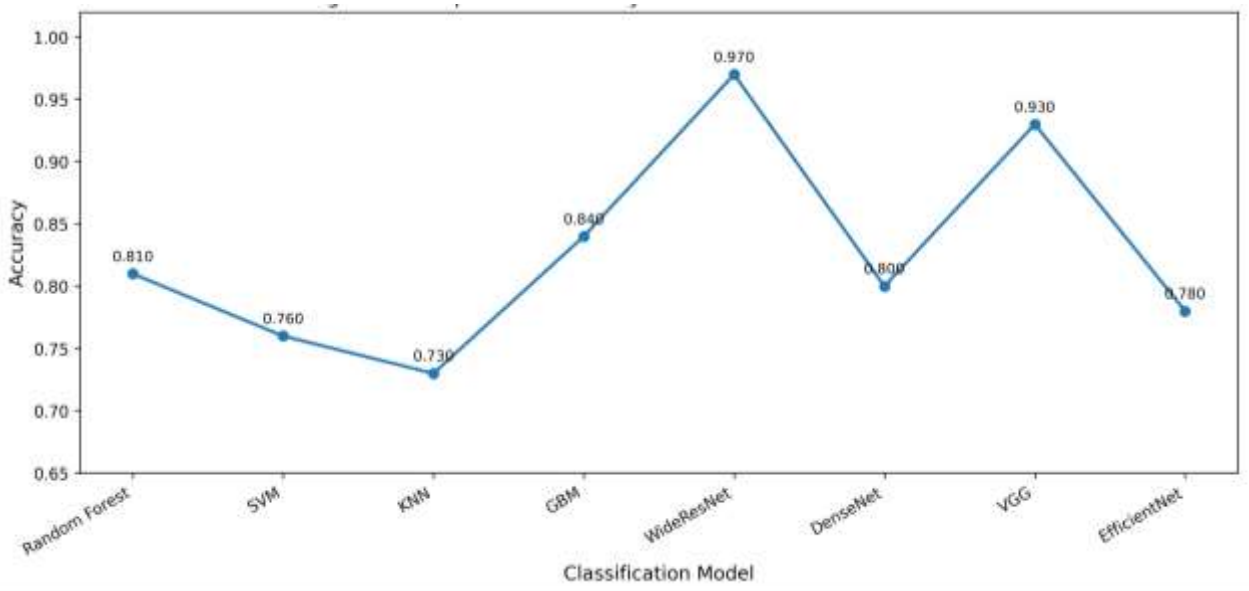

Figure 2: Comparative accuracy curve across all evaluated classification models on DUAE-reduced HNSCC gene expression data.

3.4 ROC-AUC comparative analysis

The comparative ROC-AUC performance across all models is presented in Figure 3. Consistent with the accuracy-based findings, WideResNet demonstrated the highest discriminative capability, achieving a ROC-AUC of 0.990, followed by VGG with a ROC-AUC of 0.960. Among the traditional machine learning baselines, GBM and Random Forest showed relatively strong AUC values of 0.880 and 0.860, respectively, whereas SVM and KNN achieved lower AUC values of 0.820 and 0.790. The ROC-AUC pattern reinforces the conclusion that the DUAE-reduced latent space preserved meaningful class-separating information. Importantly, the ranking observed in ROC-AUC closely mirrors the ranking observed in accuracy, suggesting that the stronger-performing models were not only accurate at a fixed threshold but also maintained superior discrimination across classification thresholds. This consistency strengthens the interpretation that WideResNet and VGG more effectively exploited the reduced gene expression representation.

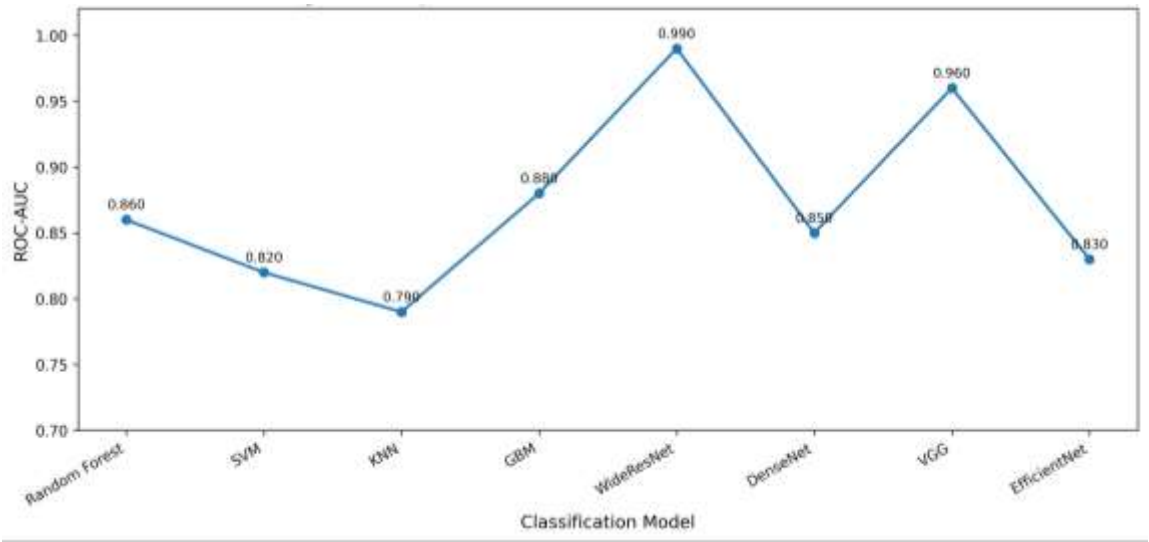

Figure 3: Comparative ROC-AUC summary across all evaluated classification models on DUAE-reduced HNSCC gene expression data.

3.5 Multi-metric performance comparison

A consolidated multi-metric comparison is shown in Figure 4, where accuracy, ROC-AUC, precision, recall, and F1-score are presented simultaneously for all classifiers. This visualization confirms that WideResNet consistently achieved the highest values across all evaluation metrics, indicating not only strong overall predictive accuracy but also a balanced performance profile with high sensitivity and precision. VGG also maintained consistently high values across all five metrics, confirming its role as the second-best model in the study.

Among the traditional machine learning baselines, GBM exhibited the strongest overall balance, outperforming Random Forest, SVM, and KNN across most metrics. Random Forest remained competitive but showed slightly lower recall and F1-score than GBM. In contrast, SVM and KNN demonstrated uniformly lower scores, reflecting weaker sensitivity and reduced overall classification robustness. Similarly, DenseNet and EfficientNet showed moderate but consistently lower values than the top two deep learning models.

The grouped multi-metric comparison is particularly important because it demonstrates that the superiority of WideResNet was not limited to a single metric. Instead, its advantage was maintained across discrimination (ROC-AUC), correctness (accuracy), positive predictive value (precision), sensitivity (recall), and harmonic balance (F1-score). This makes WideResNet the most stable and reliable model under the evaluated setting.

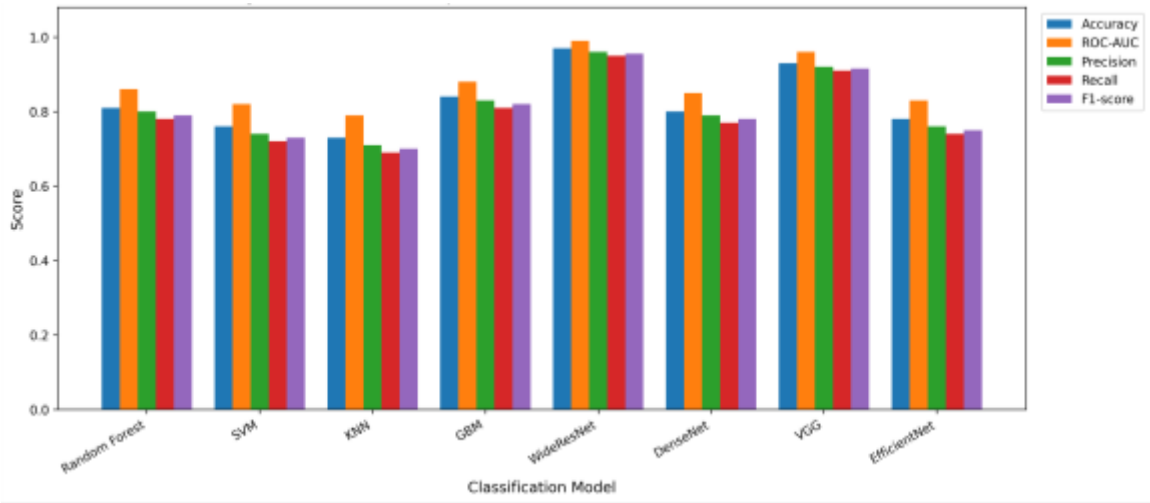

Figure 4: Multi-metric comparative performance of all evaluated classifiers, including accuracy, ROC-AUC, precision, recall, and F1-score.

3.6 Radar profile of classifier behavior

The overall classifier performance profiles are summarized in Figure 5 using a radar plot. The broader and more uniformly expanded polygons of WideResNet and VGG indicate a more balanced performance across all evaluation dimensions. In contrast, the narrower polygons of KNN, SVM, and EfficientNet reflect comparatively weaker and less consistent classification behavior.

The radar plot provides an intuitive representation of model stability across metrics. While some models may perform reasonably well on one or two criteria, a strong classifier in biomedical applications should ideally demonstrate balanced behavior across all key performance measures. In this regard, WideResNet exhibited the most comprehensive and stable performance profile, followed by VGG. The moderate but comparatively smaller profiles of GBM and Random Forest suggest that they remain viable and competitive traditional baselines, particularly when lower model complexity or faster computation is desirable.

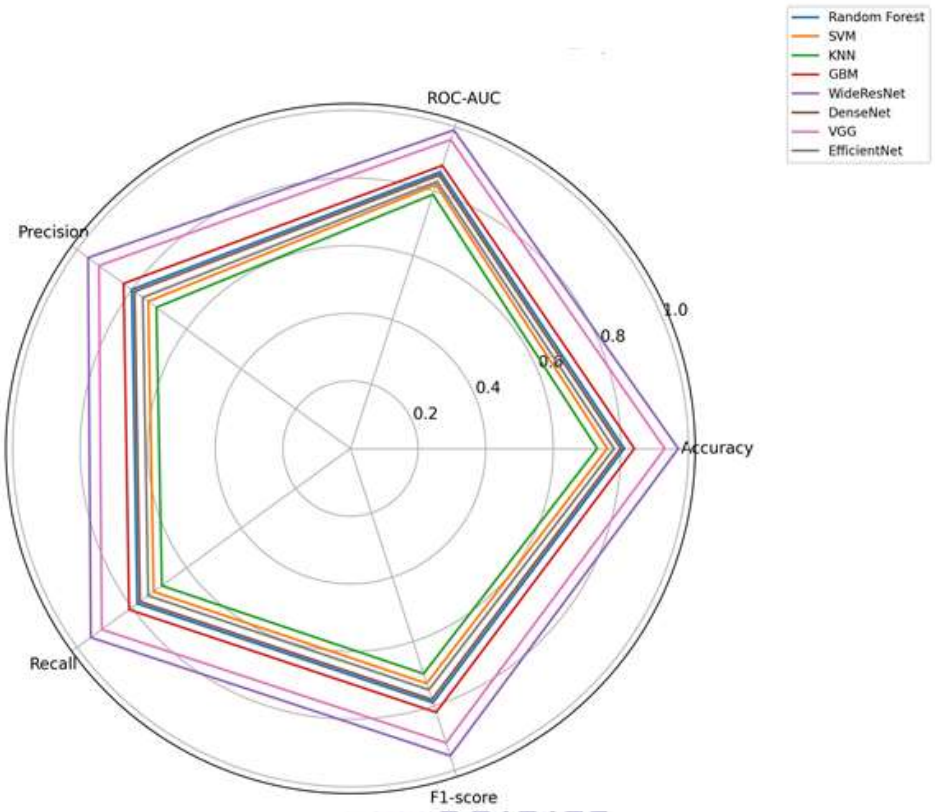

Figure 5: Radar plot of classifier performance profiles across all evaluation metrics.

4. Discussion

The present study evaluated the comparative performance of traditional machine learning and deep learning classifiers on DUAE-reduced high-dimensional HNSCC gene expression data, with the objective of determining how effectively different classifier families could exploit the compressed transcriptomic representation. The results consistently showed that while the DUAE-based dimensionality reduction preserved substantial discriminatory information, the extent to which this information translated into predictive performance depended strongly on the downstream classifier architecture.

Across all evaluation metrics, WideResNet emerged as the strongest-performing model, achieving the highest values for accuracy (0.970), ROC-AUC (0.990), precision (0.960), recall (0.950), and F1-score (0.955). This performance pattern was consistently reflected across the tabulated results and all comparative visualizations, including the accuracy curve, ROC-AUC summary, multi-metric bar plot, and radar profile. The superiority of WideResNet suggests that residual learning may be particularly well suited to DUAE-compressed transcriptomic feature spaces, where nonlinear relationships and subtle interaction structures remain embedded in the reduced representation.

A plausible explanation for the stronger performance of WideResNet lies in the role of residual connections, which facilitate more stable gradient propagation and improve feature reuse during learning. In reduced high-dimensional omics settings, the compressed latent representation may still contain complex nonlinear dependencies among genes or gene modules. Residual architectures are often better able to refine and propagate these discriminative signals without suffering from the optimization instability that can affect deeper conventional networks. In the context of the present study, this likely allowed WideResNet to learn a more

robust decision boundary than the other evaluated classifiers.

The second-best performance was observed for VGG, which also achieved a strong and well-balanced classification profile, with an accuracy of 0.930 and ROC-AUC of 0.960. Although VGG did not reach the same performance level as WideResNet, it consistently outperformed the remaining models and maintained strong precision, recall, and F1-score values. This finding indicates that the DUAE-reduced feature space retained sufficient structural regularity to support effective deep hierarchical learning, even in the absence of residual shortcuts. The relatively strong performance of VGG reinforces the broader conclusion that deep architectures can derive meaningful benefit from DUAE-based dimensionality reduction, provided that the transformed feature space remains information-rich and sufficiently organized for downstream learning.

Among the traditional machine learning models, GBM and Random Forest demonstrated the most competitive performance, with GBM outperforming Random Forest across all reported metrics. Specifically, GBM achieved an accuracy of 0.840, ROC-AUC of 0.880, and F1-score of 0.820, while Random Forest achieved an accuracy of 0.810, ROC-AUC of 0.860, and F1-score of 0.790. These findings are important because they confirm that ensemble-based machine learning methods remain highly relevant baselines in transcriptomic classification, even when compared with modern deep learning architectures. Tree-based ensemble methods are well known for their robustness to feature interactions, noise, and nonlinear boundaries, which likely contributed to their competitive performance on the DUAE-reduced latent space.

The comparatively lower performance of SVM and KNN suggests that simpler decision-boundary assumptions or distance-based neighborhood structures may be less effective in the present reduced feature setting. SVM, despite its strong theoretical suitability for high-dimensional data, achieved only moderate performance, while KNN produced the weakest overall results among the traditional baselines. One likely explanation is that the DUAE latent representation, although compact, may still preserve nonlinear class relationships that are not easily captured by a single margin-based hyperplane or by local neighborhood similarity alone. In addition, KNN can be particularly sensitive to local density variation and latent space geometry, which may limit its robustness in biologically complex data.

Among the remaining deep learning baselines, DenseNet and EfficientNet showed moderate but clearly lower performance than both WideResNet and VGG. This is an important observation because it highlights that deep learning does not guarantee superior performance in transcriptomic applications unless the architectural bias aligns with the structure of the transformed input space. DenseNet, while theoretically beneficial due to dense feature reuse, may not have provided an optimal inductive bias for the DUAE-reduced features in this study. Similarly, EfficientNet, which is highly successful in image-based scaling tasks, may not have transferred as effectively to reduced transcriptomic representations. These results underscore the importance of architecture–representation compatibility, particularly when models originally developed for image tasks are adapted to non-image biological data.

From a broader methodological perspective, the findings support the usefulness of DUAE as a dimensionality reduction framework for high-dimensional cancer transcriptomics. The consistently strong performance of the top models suggests that the DUAE preserved meaningful biological and statistical structure rather than simply compressing the data indiscriminately. In other words, the reduced latent space remained sufficiently informative for downstream classifiers to distinguish between classes with good to excellent performance. This is particularly valuable in gene expression analysis, where reducing dimensionality without losing disease-relevant signal is one of the central challenges.

At the same time, the results also emphasize that dimensionality reduction alone is not sufficient; the downstream classifier must still be carefully

selected. While the DUAE reduced the dimensional burden and likely improved the signal-to-noise ratio, different classifiers extracted very different levels of predictive value from the same reduced input. This observation is particularly important for translational bioinformatics, where methodological choices are often evaluated not only by raw performance but also by computational cost, reproducibility, interpretability, and robustness.

From an application standpoint, the results suggest a practical hierarchy for classifier selection in DUAE-based transcriptomic workflows. WideResNet appears to be the most suitable choice when maximizing predictive performance is the primary objective, whereas VGG provides a strong alternative with similarly high but slightly lower performance. If computational simplicity or interpretability is prioritized, GBM and Random Forest remain highly defensible traditional baselines and may be particularly attractive in resource-constrained settings or in studies where simpler models are preferred for reproducibility and downstream interpretation.

Despite the strong comparative trends observed in this study, several limitations should be acknowledged. First, the present work is based on a single transcriptomic dataset (TCGA HNSCC), and therefore the generalizability of the findings to other cancer types or external cohorts remains to be established. Performance on a single cohort, even when strong, does not guarantee similar behavior across different biological contexts, platforms, or preprocessing pipelines. Second, although the study employed a leakage-aware conceptual framework, future work should explicitly strengthen evaluation rigor using repeated stratified cross-validation, independent validation cohorts, or external GEO datasets to provide a more comprehensive estimate of generalizability. Third, the present graphical comparisons include model-wise ROC-AUC summaries rather than full sample-level ROC curves, meaning that future iterations of the study would benefit from reporting true ROC plots derived from prediction probabilities, as well as confusion matrices and calibration analyses. Such additions would further improve transparency and reviewer confidence.

Another important future direction involves biological interpretability. While the current study focuses on comparative predictive performance, the translational value of transcriptomic models is greatly enhanced when the reduced representation can be linked back to informative genes, pathways, or molecular mechanisms. Since the DUAE was used as a representation-learning framework, future work could investigate latent-feature attribution, decoder-based reconstruction importance, or post hoc model interpretation strategies to identify biologically relevant gene modules associated with HNSCC classification. This would strengthen the biological significance of the framework and improve its suitability for publication in more competitive bioinformatics or translational genomics venues.

Overall, the present findings demonstrate that DUAE-based dimensionality reduction can support robust downstream classification of HNSCC gene expression data, but the final performance remains highly dependent on classifier architecture. The consistent superiority of WideResNet, followed by VGG, suggests that deep residual and hierarchical convolutional feature extraction may be particularly effective for DUAE-compressed omics representations. At the same time, the competitive performance of GBM and Random Forest reinforces the continued importance of strong traditional baselines in cancer bioinformatics research.

5. Conclusion

This study presented a comparative evaluation of traditional machine learning and deep learning classifiers on DUAE-reduced high-dimensional HNSCC gene expression data. The results demonstrated that the Deep Under-complete Autoencoder (DUAE) effectively compressed the original transcriptomic space while preserving sufficient discriminative information for downstream classification. However, the ability to exploit this reduced representation varied substantially across classifier families.

Among all evaluated models, WideResNet achieved the strongest overall performance, consistently outperforming all other classifiers across accuracy, ROC-AUC, precision, recall, and F1-score. VGG emerged as the second-best model and also demonstrated strong and balanced performance across all metrics. Among the traditional machine learning baselines, GBM and Random Forest showed the most competitive results, confirming that ensemble-based approaches remain strong and scientifically relevant alternatives in reduced transcriptomic classification tasks. In contrast, SVM, KNN, DenseNet, and EfficientNet exhibited comparatively lower performance under the same DUAE-reduced input representation.

These findings highlight two key conclusions. First, DUAE-based dimensionality reduction is a promising strategy for handling high-dimensional gene expression data in cancer classification, as it can preserve meaningful class-separating structure while reducing the computational and statistical burden associated with the original feature space. Second, downstream classifier selection remains critical, as different models extract markedly different predictive value from the same reduced latent representation.

From a practical perspective, the results suggest that WideResNet should be prioritized when maximizing classification performance is the primary objective, while VGG provides a strong alternative deep learning baseline. At the same time, GBM and Random Forest remain robust and defensible baseline models, particularly in studies where model simplicity, faster training, or interpretability are desirable.

Future work should extend this framework by incorporating repeated stratified cross-validation, external validation cohorts, full sample-level ROC analysis, confusion matrices, and biomarker-level interpretability analyses. Such additions would further strengthen the methodological rigor and translational relevance of the proposed DUAE-based classification framework.

In summary, the present study supports the use of DUAE-driven representation learning combined with carefully selected downstream classifiers as a viable and effective approach for HNSCC gene expression classification, with WideResNet emerging as the most promising model under the evaluated setting.

Declarations

Ethics approval and consent to participate: This study used publicly available de-identified transcriptomic data from The Cancer Genome Atlas (TCGA).

Consent for publication: Not applicable.

Availability of data and materials: The gene expression data used in this study were obtained from The Cancer Genome Atlas (TCGA). Code and processed data can be made available by the corresponding author upon reasonable request.

Competing interests: The author declares no competing interests.

Funding: No specific external funding was received for this study.

Author contributions: A.N. conceived the study, performed the computational analysis, interpreted the results, and wrote the manuscript.

Acknowledgments: The author acknowledges the public availability of TCGA data resources and the open-source tools used for computational analysis.

References

Bayat A. Science, medicine, and the future: Bioinformatics. BMJ. 2002 Apr 27;324(7344):1018-22. doi: 10.1136/bmj.324.7344.1018. PMID: 11976246; PMCID: PMC1122955.

Lee, M. (2023). Recent Advances in Generative Adversarial Networks for Gene Expression Data: A Comprehensive Review. Mathematics, 11(14), 3055.

Thudumu, S., Branch, P., Jin, J. et al. A comprehensive survey of anomaly detection techniques for high dimensional big data. J Big Data 7, 42 (2020). https://doi.org/10.1186/s40537-020-00320-x

Fan J, Han F, Liu H. Challenges of Big Data Analysis. Natl Sci Rev. 2014 Jun;1(2):293-314. doi: 10.1093/nsr/nwt032. PMID: 25419469; PMCID: PMC4236847.

Rahnenführer, J., De Bin, R., Benner, A. et al. Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges. BMC Med 21, 182 (2023). https://doi.org/10.1186/s12916-023-02858-y

Naylor S, Chen JY. Unraveling human complexity and disease with systems biology and personalized medicine. Per Med. 2010 May;7(3):275-289. doi: 10.2217/pme.10.16. PMID: 20577569; PMCID: PMC2888109.

Sebastian, Anu Maria, and David Peter. "Artificial intelligence in cancer research: trends, challenges and future directions." Life 12.12 (2022): 1991.

Garcia Santa Cruz, B., Husch, A., & Hertel, F. (2023). Machine learning models for diagnosis and prognosis of Parkinson's disease using brain imaging: General overview, main challenges and future directions. Frontiers in Aging Neuroscience, 15, 1216163.

Historically, a myriad of machine learning models, including Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), have been employed to navigate the labyrinth of genetic data.

Singh RK, Sivabalakrishnan M. Feature selection of gene expression data for cancer classification: a review. In: Procedia Computer Science. 2015, pp. 52–7.

Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422.

Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinform 2006; 7: 1–13.

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review. Journal of Data Analysis and Information Processing, 8(4), 341-357.

Lee, L. H., Wan, C. H., Yong, T. F., & Kok, H. M. (2010). A review of nearest neighbor-support vector machines hybrid classification models. Journal of Applied Sciences, 10(17), 1841-1858.

Danilo Bzdok, Martin Krzywinski, Naomi Altman. Machine learning: Supervised methods, SVM and kNN. Nature Methods, 2018, pp.1-6. ffhal-01657491f

Chen, RC., Dewi, C., Huang, SW. et al. Selecting critical features for data classification based on machine learning methods. J Big Data 7, 52 (2020). https://doi.org/10.1186/s40537-020-00327-4

Islam, M. M., Iqbal, H., Haque, M. R., & Hasan, M. K. (2017, December). Prediction of breast cancer using support vector machine and K-Nearest neighbors. In 2017 IEEE region 10 humanitarian technology conference (R10-HTC) (pp. 226-229). IEEE.

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. arXiv preprint arXiv:1605.07146.

Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

Alves, J. H., & de Oliveira, L. F. (2020, July). Optimizing neural architecture search using limited gpu time in a dynamic search space: a gene expression programming approach. In 2020 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE.

Chen P, Li Z, Hong Z, Zheng H, Zeng R. Tumor type classification and candidate cancer-specific biomarkers discovery via semi-supervised learning. Biophys Rep. 2023 Apr 30;9(2):57-66. doi: 10.52601/bpr.2023.230005. PMID: 37753058; PMCID: PMC10518520.

Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN COMPUT. SCI. 2, 420 (2021). https://doi.org/10.1007/s42979-021-00815-1

Ahmed, S. F., Alam, M. S. B., Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., ... & Gandomi, A. H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artificial Intelligence Review, 1-97.

Spratling, M. W. (2023). Comprehensive Assessment of the Performance of Deep Learning Classifiers Reveals a Surprising Lack of Robustness. arXiv preprint arXiv:2308.04137.

Kosinski M (2022). RTCGA: The Cancer Genome Atlas Data Integration. R package version 1.28.0, https://rtcga.github.io/RTCGA.

Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68-77. doi: 10.5114/wo.2014.47136. PMID: 25691825; PMCID: PMC4322527.