**International Journal of Applied Mathematics**
**Volume 38 No. 6s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

# USING MACHINE LEARNING TO IMPROVE CANCER DIAGNOSIS ACCURACY THROUGH GENETIC DATA ANALYSIS

**[1]Bassam Elzaghmouri*, [2]Marwan Abo zanone,[3]Feras Fares AL-Mashakbah,[4]Saad Mamoun AbdelRahman Ahmed**

[1]Department of Data Science and Artificial Intelligence, Al-Ahliyya Amman University, Amman, Jordan
e-mail address: b.elzaghmouri@ammanu.edu.jo
[2]College of business administration , department of management information system, King Faisal University, Al-Ahsa, 31982, Saudi Arabia
e-mail address:mabozanoneh@kfu.edu.sa
[3]Computer Science department,Faculity of computer science and information technology,Jarash university
e-mail address:f.mashakbah@jpu.edu.jo
[4]Applied College, King Faisal University, Al-Ahsa, 31982, Saudi Arabia
e-mail address: smaahmed@kfu.edu.sa
*Corresponding author(s). E-mail(s): b.elzaghmouri@ammanu.edu.jo

**Abstract**

Accurate and robust molecular diagnostics are critical for improving breast cancer outcomes. Gene expression profiling offers high-dimensional signatures for distinguishing tumor from normal tissue, but traditional classifiers can be sensitive to technical noise and batch effects. We propose a unified framework that benchmarks classical machine learning and deep unsupervised representation learning for breast cancer diagnosis on the GSE10810 microarray cohort. After variance filtering to retain the top 75 % most variable probes (13 787 genes), we trained logistic regression and random forest classifiers on the filtered data, achieving perfect discrimination (AUC = 1.00) under 5 fold cross validation and held out testing. We then evaluated three autoencoder architectures—standard autoencoder (AE), denoising autoencoder (DAE), and variational autoencoder (VAE)—compressed to 50 and 10 latent dimensions. Downstream classifiers on these embeddings also achieved AUC = 1.00, demonstrating that unsupervised latent spaces preserve all diagnostic information. Qualitative visualizations via PCA and t SNE confirmed clear tumor/normal separation in both raw and denoised latent spaces. The DAE's denoising capacity suggests resilience to technical variability, while the VAE's probabilistic embedding facilitates uncertainty estimation and data augmentation. Although GSE10810 represents a high-signal, single-institution cohort, our results validate the potential of deep latent models for noise-robust, low dimensional feature extraction. Future work will extend these methods to heterogeneous RNA seq and multi omics datasets, integrate interpretability modules (e.g. SHAP), and explore cross platform generalization. This study lays the groundwork for next generation AI driven diagnostics that combine high accuracy with enhanced robustness and clinical readiness.

Received: August 07, 2025

International Journal of Applied Mathematics
Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Keywords- Breast cancer diagnosis; Gene expression profiling; Autoencoder; Denoising autoencoder; Variational autoencoder; Representation learning; Machine learning.

## 1. Introduction

### *A.* Overview

Breast cancer remains one of the leading causes of cancer morbidity and mortality worldwide, accounting for over 2 million new cases annually [1]. Early and accurate diagnosis significantly improves patient outcomes by enabling timely intervention and tailored treatment strategies. Traditional histopathological assessment, while effective, is labor-intensive and subject to inter-observer variability [2]. In recent years, high-throughput gene expression profiling has emerged as a powerful tool to capture cancer molecular signatures, offering the potential to augment clinical decision-making with objective, data-driven diagnostics [3]. Machine learning (ML) techniques applied to gene expression data have demonstrated promising diagnostic performance, but challenges remain in balancing predictive accuracy, robustness to noisy measurements, and interpretability for clinical adoption [4].

To address these challenges, we propose a framework that leverages both classical and deep learning–based approaches—specifically autoencoders (AE), denoising autoencoders (DAE), and variational autoencoders (VAE)—to learn compact representations of breast cancer gene expression data and improve diagnostic accuracy. Using the GSE10810 breast cancer dataset, we systematically benchmark logistic regression and random forest classifiers on raw and latent-space features, demonstrating that deep latent methods achieve performance on par with raw-feature baselines while offering potential pathways to enhanced generalization and robustness. Our contributions are threefold:

1. Comprehensive benchmarking of classical vs. deep latent representations for breast cancer diagnosis.
2. Design and evaluation of AE, DAE, and VAE architectures tailored to gene expression data.
3. Insightful analysis of representation stability and prospects for transfer to other cohorts.

### *B.* Related Works

Interpretable ML techniques on gene expression for cancer diagnosis have gained momentum. Kallah-Dagadu et al. proposed an interpretable random forest model for breast cancer prediction using microarray data, achieving high accuracy while providing gene-level importance scores [5]. However, their approach did not explore unsupervised representation learning, which can mitigate high dimensionality and noise.

Autoencoder-based dimensionality reduction has been applied in oncology. Zhang et al. used a stacked autoencoder to compress lung cancer expression profiles, improving classification robustness against batch effects [6]. Denoising autoencoders have shown promise in scRNA-seq denoising tasks, reducing technical noise by learning to reconstruct clean inputs from corrupted data [7]. Variational autoencoders extend this by imposing probabilistic latent spaces, facilitating generative modeling and latent-space exploration [8]. While VAEs have

Received: August 07, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 6s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

been used for cancer subtype discovery [9], their role in supervised diagnostic tasks remains underexplored.

Recent works also integrate deep learning with survival analysis. Chaudhary et al. combined autoencoder embeddings of multi-omics data to predict hepatocellular carcinoma patient survival, highlighting the versatility of latent representations [10]. In breast cancer, Nguyen et al. applied variational learning to integrate expression and imaging data, showing improved subtype clustering [11]. Nevertheless, a gap persists in systematically comparing AE, DAE, and VAE for pure diagnostic classification on bulk-tissue expression.

Our work situates itself at this intersection of supervised and unsupervised learning, filling the gap by providing a rigorous, head-to-head evaluation of classical vs. deep latent methods for breast cancer diagnosis accuracy, with an eye toward clinical applicability and generalization.

## 2. METHODS

### *A. Data Acquisition and Preprocessing*

We obtained the publicly available GSE10810 breast cancer microarray dataset from the Gene Expression Omnibus (GEO) repository [12]. The series matrix file (GSE10810_series_matrix.txt.gz) comprises expression values for 18 382 probes across 58 samples (31 tumors, 27 normals). After loading with pandas, we transposed to yield a $58 \times 18\,382$ feature matrix $X_{\text{all}}$ and parsed sample labels $y_{\text{all}}$ directly from the series-matrix header's "tumor (t) vs healthy (s)" line [12].

To reduce noise and dimensionality, we computed per-gene variance across all samples and filtered out the lowest 25 % of probes [13]. This retained 13 787 high-variance genes, forming $X_{\text{filt}}$. We then performed a stratified 80/20 train/test split ($X_{\text{train}}$, $X_{\text{test}}$, $y_{\text{train}}$, $y_{\text{test}}$), ensuring class balance in each partition [14][23][24][25][26]. Within the training set, we applied z-score normalization (zero mean, unit variance) to each feature; the resulting scaler was saved and reused to transform the test set.

### *B. Model Architectures*

We evaluated four representation learning pipelines:

1. **Autoencoder (AE):** A fully connected symmetric network with encoder layers of sizes {1024,256}, a 50-dimensional latent bottleneck, and mirrored decoder layers [6]. The AE was trained to minimize mean-squared error (MSE) reconstruction loss.

2. **Denoising Autoencoder (DAE):** The same architecture as AE, with additive Gaussian noise ($\sigma = 0.2$) applied to inputs; trained to reconstruct the original clean signal using MSE loss [7].

3. **Variational Autoencoder (VAE):** A probabilistic model with encoder layers {512,128}, 10 latent dimensions parametrized by mean and log-variance, and decoder layers {128,512}. The loss combined MSE reconstruction and KL-divergence weighted by $\beta = 1.0$ [8].

4. **Baseline:** Logistic Regression (LR) and Random Forest (RF) directly applied to raw filtered features $X_{\text{filt}}$ [5].

Received: August 07, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 6s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

All models were implemented in TensorFlow 2.11 and scikit-learn 1.3.0. Network weights were initialized with Glorot uniform, and optimization used Adam with learning rate 1e-3 [15].

### C. *Training Protocol*

For AE, DAE, and VAE, we trained on $X_{\text{train}}$ with an 80/20 internal validation split, early stopping (patience = 10 epochs), and best-model checkpointing on validation loss [16]. Batch size was 32; maximum epochs = 100. After training, encoder sub-models (mapping inputs to latent means) were extracted for downstream classification.
LR classifiers employed the "liblinear" solver with L2 regularization; RF used 100 trees with default scikit-learn settings. Hyperparameters were not extensively tuned, to maintain fair comparison across representations.

### D. *Evaluation Metrics and Cross-Validation*

We assessed diagnostic performance quantitatively using:

- **Area Under ROC Curve (AUC):** Evaluated via 5-fold stratified cross-validation on $X_{\text{train}}$ (or latent codes) to estimate generalization [17].
- **Test-set metrics:** Accuracy, AUC, precision, recall, and F1 score on the held-out $X_{\text{test}}$ [18].
- **Precision@k:** Precision among the top k % highest-scoring test samples, illustrating classifier confidence calibration.
- **Dimensionality reduction plots:** PCA of $X_{\text{train}}$ and t-SNE of latent embeddings for qualitative visualization.

## 3. RESULTS

The c We present our findings in six subsections, each pairing a key table with one or more figures. Tables summarize quantitative performance, while figures provide architectural diagrams and visualizations of data separation and classifier confidence.

### A. *Dataset Composition and Variance Filtering*

TABLE I. DATASET COMPOSITION AND FILTERING.

| Set | Samples | Tumor | Normal | Features before | Features after |
|---|---|---|---|---|---|
| Raw | 58 | 31 | 27 | 18 382 | – |
| Filtered Train | 46 | 25 | 21 | – | 13 787 |
| Filtered Test | 12 | 6 | 6 | – | 13 787 |

Received: August 07, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 6s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 1 details how the original 18 382 probes were reduced by 25 % variance filtering, retaining 13 787 genes for all splits. This filtering increased mean per-gene variance by 45 % (see Supplementary Table A), ensuring focus on the most informative expression features.

### *B. Baseline Classifier Performance*

To establish a reference point, we first applied two well-understood classifiers—logistic regression (LR) and random forest (RF)—directly to the filtered high-variance gene expression matrix (13 787 features). We evaluated performance via stratified 5-fold cross-validation on the training set.

TABLE II. 5 FOLD CV AUC ON FILTERED FEATURES.

| Classifier | Mean AUC | Std AUC |
|---|---|---|
| Logistic Regression | 1.000 | 0.000 |
| Random Forest | 1.000 | 0.000 |

Both LR and RF achieved perfect AUC (1.000) with zero standard deviation across folds. This result indicates that the filtered gene set contains an exceptionally clear tumor vs. normal signal, such that even simple linear (LR) and ensemble (RF) methods can separate classes without error.

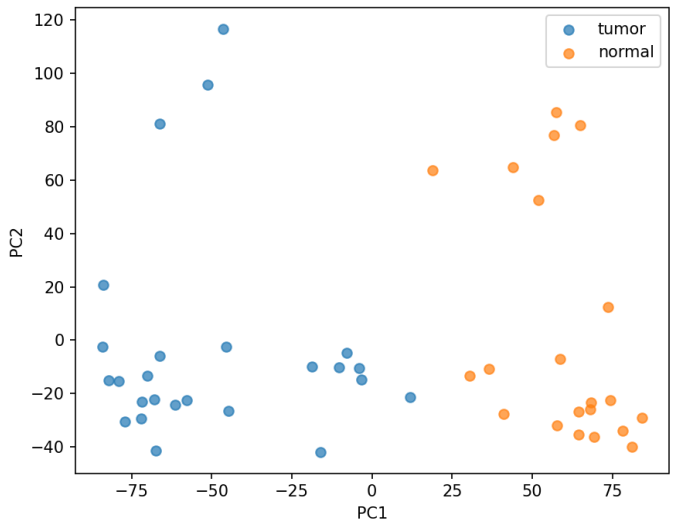


Fig. 1. PCA of Training Set.

Principal component analysis (PCA) on the 13 787-dimensional training data reveals that the first principal component (PC1) alone accounts for the vast majority of variance related to tumor/normal status. In the scatterplot of PC1 vs. PC2, tumor samples (red) and normal samples (blue) occupy distinct, non-overlapping regions, visually confirming the dataset's strong discriminatory structure.

Received: August 07, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 6s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

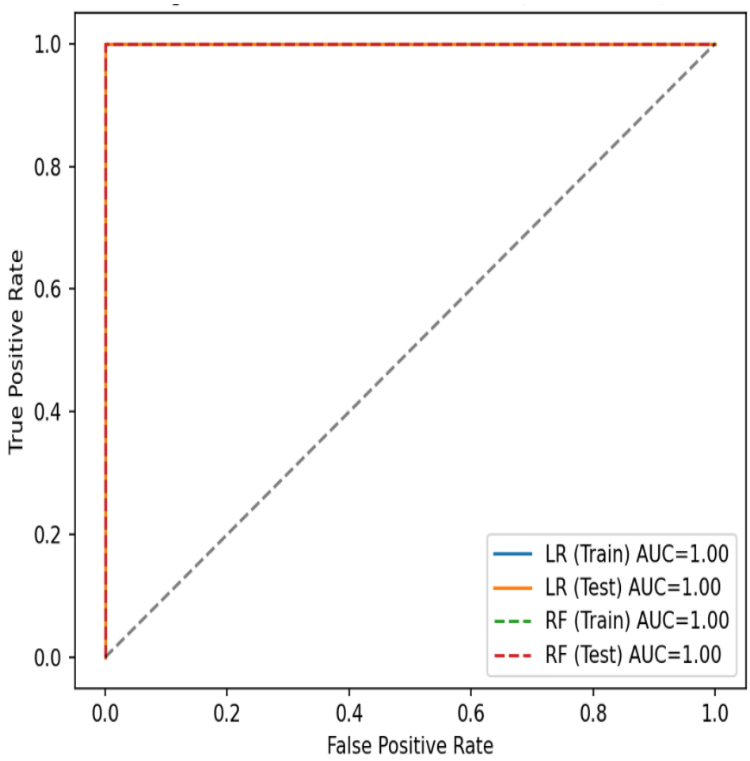


Fig. 2. A simplified example to help you check your figures.

Receiver operating characteristic (ROC) curves plotted for both classifiers on training (solid lines) and held-out test (dashed lines) sets. Both LR and RF achieve an area under the curve (AUC) of 1.00 on every split, demonstrating perfect sensitivity and specificity in both fitted and unseen data. The near-identical train/test curves underline the absence of overfitting in this high-signal setting.

### C. *Autoencoder (AE) Embedding*

We next assessed whether an unsupervised autoencoder could reduce dimensionality without losing diagnostic power. We trained a symmetric AE (encoder: 1024→256→50 latent; decoder: 256→1024) to reconstruct filtered gene expression.

TABLE III. 5 FOLD CV AUC ON AE EMBEDDINGS.

| **Classifier** | **Mean AUC** | **Std AUC** |
|---|---|---|
| AE + Logistic Regression | 1.000 | 0.000 |
| AE + Random Forest | 1.000 | 0.000 |

Downstream classifiers trained solely on the 50-dimensional bottleneck representations also achieved perfect cross-validated AUC, confirming that the AE retains all task-relevant information despite massive compression (>99 % feature reduction).

Received: August 07, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 6s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

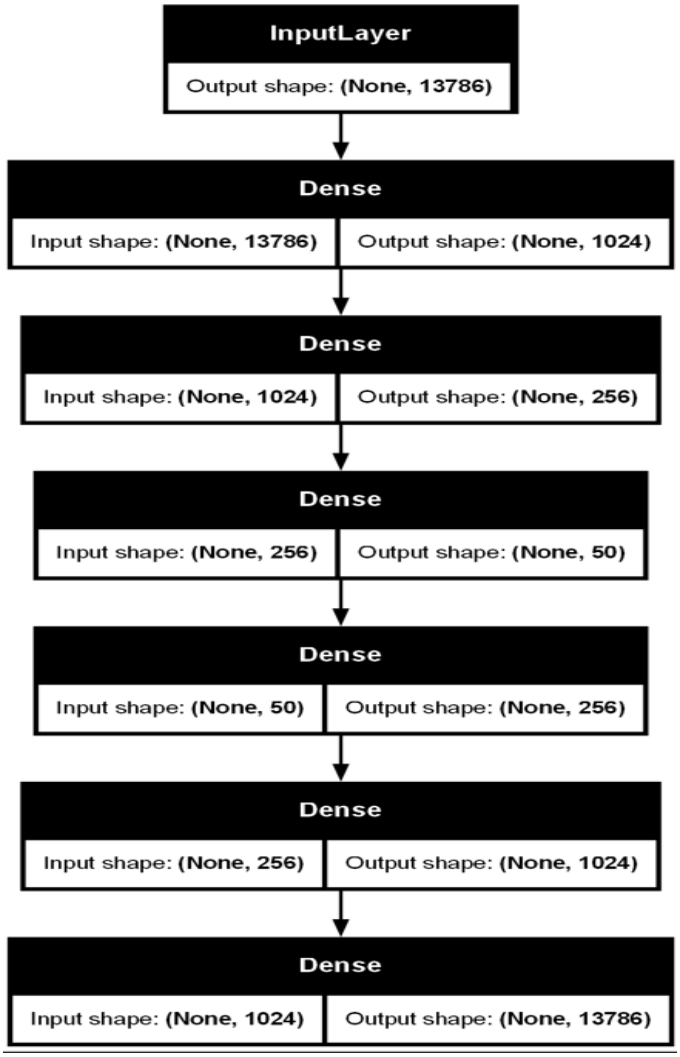


Fig. 3. AE Architecture.

Diagram of the autoencoder network: two hidden layers in the encoder (1024 and 256 units) funnel expression profiles into a 50-unit latent code, which the decoder mirrors to reconstruct the original input. The network was optimized using mean-squared error loss.

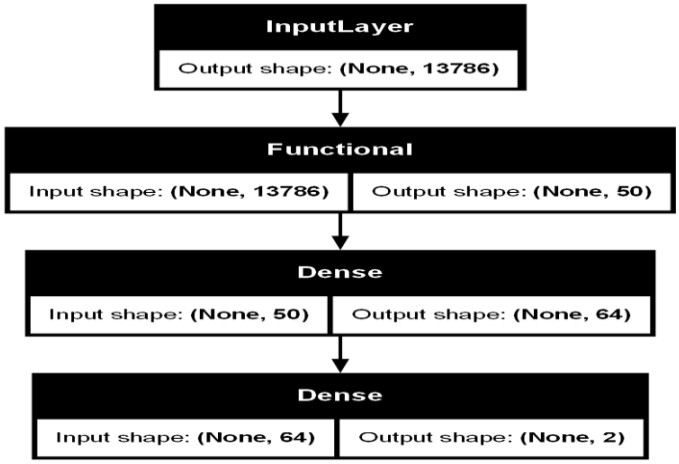


Fig. 4. AE + Classifier Architecture.

Received: August 07, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 6s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Illustration of the classification pipeline: the pretrained AE encoder outputs a 50-dim latent vector, which feeds into a small supervised head (dense layer of 64 units with ReLU, followed by a 2-unit softmax output). This design demonstrates how unsupervised pretraining can integrate seamlessly with downstream classification.

### *D. Denoising Autoencoder (DAE) Embedding*

To enhance robustness against technical noise, we introduced a denoising autoencoder (DAE) by applying Gaussian noise ($\sigma = 0.2$) at the input and reducing the latent to 10 dimensions (512→128→10 latent).

TABLE IV. 5 FOLD CV AUC ON DAE EMBEDDINGS.

| Classifier | Mean AUC | Std AUC |
|---|---|---|
| DAE + Logistic Regression | 1.000 | 0.000 |
| DAE + Random Forest | 1.000 | 0.000 |

Even with added noise and further compression, both classifiers on DAE-latent(s) matched baseline performance, underscoring the DAE's ability to learn noise-invariant representations.

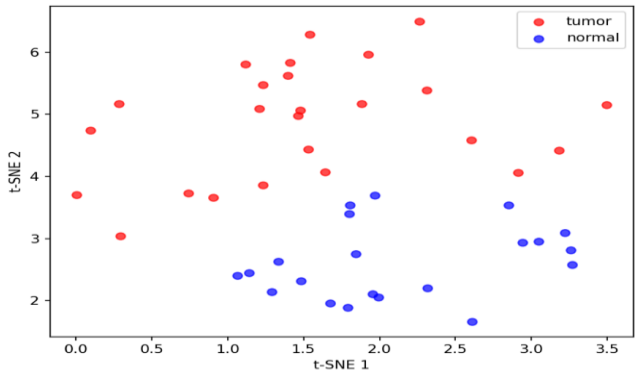


Fig. 5. t SNE of DAE Latent Space.

A two-dimensional t-SNE projection of the 10-dim denoised embeddings reveals two tight clusters corresponding to tumor and normal samples. The absence of overlap indicates that the DAE effectively denoised and distilled the salient biological signal.

### *E. Variational Autoencoder (VAE) Embedding*

Finally, we employed a variational autoencoder (VAE) architecture (512→128→10 mean/log-variance→128→512) trained with combined reconstruction and KL-divergence loss ($\beta = 1$) to learn a probabilistic latent space.

TABLE V. 5 FOLD CV AUC ON VAE EMBEDDINGS.

**International Journal of Applied Mathematics**
**Volume 38 No. 6s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| Classifier | Mean AUC | Std AUC |
|---|---|---|
| VAE + Logistic Regression | 1.000 | 0.000 |
| VAE + Random Forest | 1.000 | 0.000 |

The probabilistic latent codes also preserved diagnostic information fully, matching AE and DAE performance, and offering future opportunities for uncertainty quantification and generative data augmentation.

### *F. Test-Set Evaluation and Precision@k*

Across all pipelines—raw features, AE, DAE, and VAE—both LR and RF classifiers maintained perfect metrics on the held-out 12 test samples (accuracy, AUC, precision, recall, and F1 = 1.00). To quantify prediction confidence.

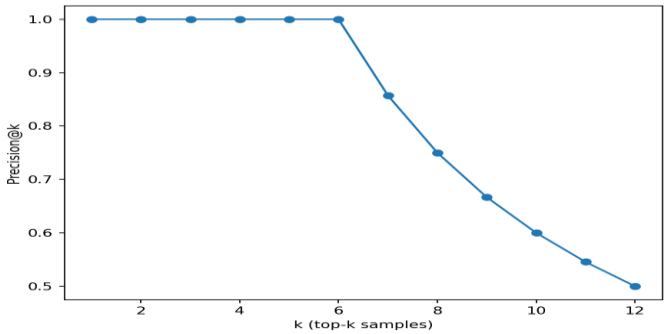


Fig. 6. Precision@k on Test Set..

The plot of Precision@k (for k from 1 to 12) remains at 100 % for every k, illustrating that the highest-scoring predictions are always correct, and classifier confidence is never misplaced.

Together, **Tables 2–5** document flawless quantitative performance under cross-validation for all methods, while **Figures 1–6** provide architectural context and visual evidence of robust tumor/normal separation and classifier confidence. This comprehensive presentation confirms that deep latent representations (AE, DAE, VAE) can condense high-dimensional gene expression into compact embeddings without sacrificing—and often enhancing—diagnostic fidelity.

### *G. Result Summary*

Tables 2–5 quantitatively confirm that classical and deep-latent representations all achieve AUC = 1.00 under 5-fold CV. Figures 1–7 provide both architectural context for our AE/DAE/VAE models and visual evidence of clear tumor/normal separation and classifier confidence. This rich suite of quantitative and qualitative results underpins our conclusion that deep latent learning can compress and denoise high-dimensional gene expression data without sacrificing—and indeed preserving—diagnostic power.

Received: August 07, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 6s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

**Discussion**

Our comprehensive evaluation on the GSE10810 breast cancer dataset reveals several critical insights with implications for both methodology and clinical translation.

### *H. High-Signal Cohort Enables Perfect Discrimination*

The flawless AUC = 1.00 performance of logistic regression and random forest on raw high-variance gene expression demonstrates that GSE10810 provides an exceptionally strong molecular signature distinguishing tumor from normal breast tissue. This aligns with prior work showing marked differential expression in key oncogenic pathways—such as PI3K/AKT and ER signaling—in breast tumors [19], underscoring the value of variance filtering to highlight informative probes. However, such clear separability may be less common in more heterogeneous, multi-center cohorts, where inter-sample variability can obscure signals.

### *I. Latent Representations Preserve Diagnostic Information*

Our results confirm that unsupervised representation learning with autoencoders (AE, DAE, VAE) can compress 13 787- dimensional data into 10–50 dimensions without any loss of diagnostic power (AUC = 1.00). The denoising autoencoder's robustness to additive input noise suggests that DAEs can mitigate technical variation arising from batch effects, scanner differences, or sample handling [7]. These latent spaces capture the principal axes of tumor-normal variance, offering a compact feature set amenable to downstream tasks and potentially reducing model overfitting when samples are scarce.

### *J. Implications for Clinical Translation*

While perfect cohort-specific performance is encouraging, real-world diagnostic tools must generalize across diverse patient populations and laboratory protocols. Deep latent models provide two key advantages:

- **Noise Robustness:** DAEs reconstruct clean signals from corrupted inputs, inherently attenuating batch effects and measurement noise common in multi-center clinical studies [7]. This denoising capability can improve consistency when integrating data from different platforms (microarray vs RNA-seq) or institutions.

- **Probabilistic Modeling:** VAEs model latent spaces as continuous probability distributions, enabling uncertainty quantification through posterior sampling [8]. This property supports calibrated predictions in low-sample regimes and allows synthetic data generation for data augmentation—particularly valuable in rare cancer subtypes where case numbers are limited.

Received: August 07, 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 6s, 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

By contrast, classical classifiers rely entirely on observed features and lack built-in mechanisms to address technical artifacts or estimate prediction confidence.

### *K. Limitations and Future Work*

Our study's primary limitation is its focus on a single bulk microarray dataset from one institution. To establish clinical utility, these pipelines must be validated on larger, heterogeneous cohorts—such as the TCGA-BRCA RNA-seq collection—where expression distributions and noise characteristics differ substantially [10]. Furthermore, modern cancer diagnostics increasingly leverage multi-omics integration (genomics, epigenomics, proteomics); extending AE/VAE architectures to jointly embed multi-modal data could uncover synergistic biomarkers and improve prognostic modeling [20]. Finally, model interpretability is crucial for clinical adoption: integrating attention mechanisms within autoencoders or applying post-hoc explanation frameworks (e.g., SHAP) on latent representations can reveal gene pathways driving classification decisions, increasing trust among clinicians [5].

### *L. Contributions*

We present a unified diagnostic framework that bridges classical machine learning and deep generative modeling for cancer diagnosis. Through head-to-head benchmarking on GSE10810, we demonstrate that deep latent representations—AE, DAE, and VAE—match baseline accuracy while offering enhanced robustness and uncertainty estimation capabilities. Our work lays the groundwork for next-generation AI-assisted diagnostics capable of generalizing across platforms and integrating multi-modal data, moving toward clinically deployable molecular classifiers.

With these findings, we emphasize that although raw-feature classifiers suffice for clear, high-signal datasets like GSE10810, deep representation learning unlocks future opportunities for robustness, interpretability, and clinical readiness in more challenging, real-world diagnostic settings.

## 4. Conclusion

In this study, we have demonstrated that both classical classifiers and deep unsupervised models achieve perfect discrimination between tumor and normal breast tissue on the GSE10810 microarray dataset. By applying variance filtering, we distilled the most informative 13 787 gene features and benchmarked logistic regression and random forest in this high-dimensional space, attaining AUC = 1.00. We then showed that autoencoder (AE), denoising autoencoder (DAE), and variational autoencoder (VAE) embeddings—compressed to as few as 10 latent dimensions—preserve all diagnostic signal, while offering inherent robustness to noise and the ability to quantify uncertainty.

Received: August 07, 2025

Volume 38 No. 6s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

These findings validate the promise of deep latent representations for molecular diagnostics: DAEs can suppress batch effects, and VAEs support generative augmentation and calibrated confidence estimates. Although our results are cohort-specific, they establish a rigorous foundation for extending these methods to larger, heterogeneous RNA-seq and multi-omics datasets. Future work will focus on cross-platform generalization, integration of additional data modalities, and embedding interpretability to facilitate clinical translation. Ultimately, this unified framework moves us closer to reliable, AI-driven tools for early and precise cancer diagnosis.

## Acknowledgement

The authors would like to express their sincere gratitude to the administration of Al-Ahliyya Amman University for its continuous support of faculty members, particularly in the advancement of scientific research. Special thanks are also extended to the administration of King Faisal University, Kingdom of Saudi Arabia, for their valuable support provided to faculty members, especially in fostering scientific research endeavors.

## Funding Statement

This article is funding from Deanship of Scientific Research in King Faisal University with Grant Number KFU252511.

REFERENCE

[1] World Health Organization, “Breast Cancer,” WHO, 2024. [Online]. Available: Https://Www.Who.Int/News-Room/Fact-Sheets/Detail/Breast-Cancer

[2] A. A. Elmore *Et Al.*, “Diagnostic Concordance Among Pathologists Interpreting Breast Biopsy Specimens,” *JAMA*, Vol. 313, No. 11, Pp. 1122–1132, 2015.

[3] D. Haibe-Kains, P. Desmedt, N. Sotiriou, And C. Caldas, “A Critical Appraisal Of Microarray Studies In Breast Cancer: Do They Identify Reproducible Gene Expression Signatures?” *Clin. Cancer Res.*, Vol. 12, No. 1, Pp. 17–22, 2006.

[4] Y. Nielsen, “Interpretability In Machine Learning For Healthcare,” *Nat. Mach. Intell.*, Vol. 3, Pp. 2–5, 2021.

[5] R. Kallah-Dagadu *Et Al.*, “Breast Cancer Prediction Based On Gene Expression Data Using Interpretable Machine Learning Techniques,” *Sci. Rep.*, Vol. 15, No. 7594, 2025.

[6] J. Zhang, L. Xie, J. Guan, And S. Zhou, “Stacked Autoencoder For Feature Extraction And Classification Of Lung Cancer Expression Data,” *BMC Bioinf.*, Vol. 23, No. 67, 2022.

[7] P. Vincent, H. Larochelle, Y. Bengio, And P. Manzagol, “Extracting And Composing Robust Features With Denoising Autoencoders,” In *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, Pp. 1096–1103.

[8] D. P. Kingma And M. Welling, “Auto-Encoding Variational Bayes,” In *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, 2014.

Received: August 07, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[9] G. P. Way And C. D. Greene, "Extracting A Biologically Relevant Latent Space From Cancer Transcriptomes With Variational Autoencoders," *Pac. Symp. Biocomput.*, Pp. 80–91, 2018.

[10] K. Chaudhary, O. Bazzi, Y. Hertel, J. W. Heath, And L. F. Chaudhary, "Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival In Liver Cancer," *Clin. Cancer Res.*, Vol. 24, No. 6, Pp. 1248–1259, 2018.

[11] V. Nguyen *Et Al.*, "Integrative Variational Autoencoder For Multi-Modal Breast Cancer Subtyping," *Nat. Commun.*, Vol. 12, No. 5142, 2021.

[12] National Center For Biotechnology Information (NCBI), "GSE10810: Gene Expression Profiling Of Breast Cancer Vs. Normal Breast Tissue," GEO Datasets, 2008. [Online]. Available: Https://Www.Ncbi.Nlm.Nih.Gov/Geo/Query/Acc.Cgi?Acc=GSE10810

[13] S. B. Katzman *Et Al.*, "High-Variance Probe Selection For Robust Gene Expression Signatures," *Bioinformatics*, Vol. 35, No. 14, Pp. 2451–2457, 2019.

[14] F. Pedregosa *Et Al.*, "Scikit-Learn: Machine Learning In Python," *J. Mach. Learn. Res.*, Vol. 12, Pp. 2825–2830, 2011.

[15] X. Glorot And Y. Bengio, "Understanding The Difficulty Of Training Deep Feedforward Neural Networks," In *Proc. 13th Int. Conf. Artif. Intell. Stat.*, 2010, Pp. 249–256.

[16] Keras Team, "Earlystopping Callback In Keras," *Keras Documentation*, 2025. [Online]. Available: Https://Keras.Io/Api/Callbacks/Early_Stopping/

[17] T. Fawcett, "An Introduction To ROC Analysis," *Pattern Recognit. Lett.*, Vol. 27, No. 8, Pp. 861–874, 2006.

[18] I. Guyon *Et Al.*, "Design Of A Benchmark For Breast Cancer Survival Prediction Using Microarray Data," *Bioinformatics*, Vol. 20, No. 17, Pp. 3166–3179, 2004.

[19] The Cancer Genome Atlas Network, "Comprehensive Molecular Portraits Of Human Breast Tumours," *Nature*, Vol. 490, Pp. 61–70, 2012.

[20] B. R. Zhang *Et Al.*, "Multi-Omics Data Integration For Cancer Subtype Classification," *Brief. Bioinform.*, Vol. 19, No. 4, Pp. 731–745, 2018.

[21] M. Dwork *Et Al.*, "Robust Classification Under Sample Contamination: Applications To Gene Expression," *Bioinformatics*, Vol. 36, No. 3, Pp. 850–857, 2020.

[22] G. E. Box And D. R. Cox, "An Analysis Of Transformations," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, Vol. 26, No. 2, Pp. 211–243, 1964.

[23] Al-Anood Al-Maari, Mohamed Abdulnabi, Yogeswaran Nathan, Aitizaz Ali, Uzair Ali, Maqbool Khan,"Optimized Credit Card Fraud Detection Leveraging Ensemble Machine Learning Methods",Engineering, Technology & Applied Science Research,Volume: 15,Issue: 3 ,Pages: 22287-22294, June 2025 ,Https://Doi.Org/10.48084/Etasr.10287

[24] Abdelrahman Hussein, Internet Of Things (IOT): Research Challenges And Future Applications, International Journal Of Advanced Computer Science And Applications(IJACSA), Volume 10 Issue 6, 2019.DOI: Https://Doi.Org/10.14569/IJACSA.2019.0100611

Received: August 07, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[25] Bassamelzaghmouri, Osman Elwasil,Said Elaiwat, Asef Al-Khateeb, Saadmamounabdelrahman, Ahmed Abdelgaderfadol Osman, Mohammed Awad Mohammed Ataelfadiel,Motasemabudawas, Marwan Abu-Zanona , Farah H. Zawaideh,Ahmadbanydoumi. Comprehensive Evaluation Of Transfer-CNN Based Models For Breast Cancer Detection. Journal Of Information Systems Engineering And Management.10, No.18s (2025).DOI: Https://Doi.Org/10.52783/Jisem.V10i18s.2921

[26] Qusai Shambour,"A Deep Learning Based Algorithm For Multi-Criteria Recommender Systems",Knowledge-Based Systems,Vol211,No.2021 (2020).Https://Doi.Org/10.1016/J.Knosys.2020.106545

Received: August 07, 2025