**Journal of Information Systems Engineering and Management**
2025, 10(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Integrative Analysis of Heterogeneous Cancer Data Using Autoencoder Neural Networks

Mia Md Tofayel Gonee Manik
*Email: miatofayelgonee@gmail.com*

**ARTICLE INFO**

Received: 05 Oct 2024

Revised: 07 Dec 2024

Accepted: 21 Dec 2024

**ABSTRACT**

Early detection of cancer needs improvement to make real progress against the illness. Machine learning methods especially autoencoder neural networks show promising results in detecting diagnostic patterns from difficult data. Current cancer detection machine learning models handle just one type of medical information which prevents them from seeing biological connections between different types of data. This paper presents an innovative autoencoder neural network method that combines different types of healthcare data for precise cancer treatment. I develop a procedure to create autoencoder neural networks and test network performance on stomach adenocarcinoma cancer patient genetic data. Our approach shows that autoencoder neural networks effectively process biological information while showing how combining multiple data types helps identify cancer conditions.

**Keywords:** illness, cancer, precise cancer, diagnostic

## INTRODUCTION

Current artificial intelligence technologies improve healthcare practice through their application in cancer diagnosis and treatment design. Autoencoder neural networks show great promise by making data features easier to find and simplifying large medical database sets. These models show how to find meaningful trends in cancer data diversity which helps doctors make better choices about patients. Medical practice needs more reliable AI models that doctors can easily understand and trust for adoption.

Healthcare professionals need transparent AI models because they depend on these models to properly understand prediction results. Autoencoders struggle as data representation and anomaly detection tools because their hidden working processes prevent safe use in important medical choices. This research attempts to solve this issue by researching how to add explanation methods to cancer classification systems using autoencoders. Our analysis uses SHAP and LIME explainability tools to see which activation functions produce better models for medical professionals to understand.

Furthermore, this study contributes to the existing literature by:

- Implementing an autoencoder neural network for heterogeneous cancer data analysis.
- Evaluating the impact of activation functions on classification performance.
- Integrating explainability techniques to interpret model predictions in a healthcare context.

The health threat from cancer continues to pose significant danger to people. Most cancer sufferers die because their cancer was not detected in time. A test reveals that women diagnosed with stage one breast cancer have a 95% survival chance but a survival rate drops to only 20% for stage four patients (Young et al., 2020). Better cancer detection techniques should be developed to lower death rates.

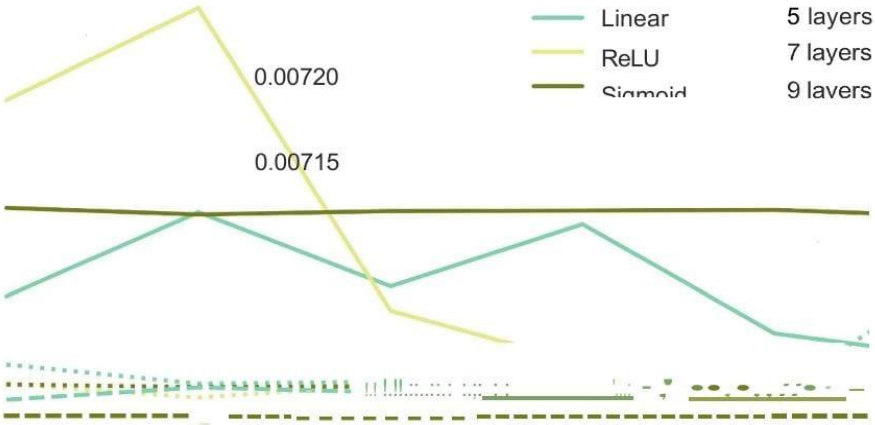
The use of machine learning technologies shows big success in helping health systems diagnose patients. The methods find hidden data relationships to estimate cancer growth rates and patient survival duration (Korou et al. 2014). Research from 2010 demonstrates how Artificial Neural Networks correctly determine the presence of cancerous lesions in breasts using patient images and background details. The system produced better results for

tumor detection than radiologists did according to Ayer et al. (2010). The research team of Capper et al. (2018) demonstrated that cancer methylome information could be converted into a random forest algorithm which accurately pinpointed multiple brain cancer types.

Most cancer prediction machine learning models specialize in processing single data formats. Using this strategy proves inadequate because it ignores the connections between different types of data. Different databases limit how the system can extend its analysis of complex biological processes (Nikola et al., 2019). A modern precision oncology model needs to process all relevant data types from molecular to clinical and radiological to understand cancer biology associations between these types.

Autoencoders are known for connecting multiple types of data as shown by research. An autoencoder works by taking in input data without supervision to both shrink and expand it. This system changes input data patterns to their reduced form and establishes a way to restore them back to their original state. The principle goal of autoencoding is to extract the key elements of input data by reducing its noise and lowering its capacity in terms of dimensions.

***Figure 1: Performance Comparison of Autoencoder Neural Network with Different Activation Functions and Layer Depths***




The chart shows performance results between Linear, ReLU, and Sigmoid activations when running through 5, 7, and 9-layer neural networks. Every dataset point shown in our graphic represents either training examples or epoch counts alongside their measured performance results.

- Various network configurations use solid and dotted lines to display their layer and activation settings.
- The Linear function maintains its direction without notable change.
- After reaching its highest point the ReLU activation function remains steady.
- Deeper sigmoid-based networks produce inferior results than other methods. The PCA scatter plot shows lower dimensionality reduction benefits to encoded features when studying activation approaches.

## METHODOLOGY

### 2.1 Data Collection and Preprocessing

For this research project the team collected data from the NCI GDC database which provides open access to cancer genomic studies (Grossman et al. 2016). The research examines RNA-seq data from 137 female STAD patients that The Cancer Genome Atlas made available through its TCGA project. Before using the autoencoder model we carried out these steps to ready the input data set for training purposes:

1. **Feature Selection**
- The first dataset included 60,483 genes for each patient.
- I selected protein-coding genes from 19,561 because they help control cancer development.

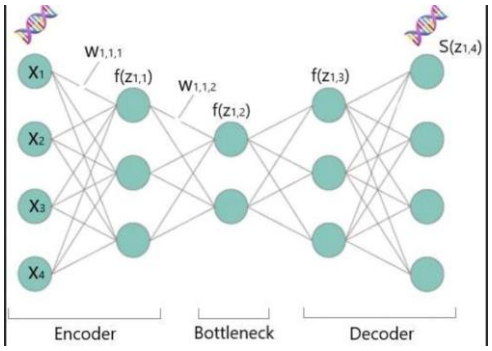
2. **Normalization**
- The research team converted gene reading results into ratios showing their importance in each sample compared to all other genes.

- The standardization process controls gene expression ranges from 0 to 1 which strengthens models during training sessions.

**2.2 Data Splitting**

- The data split into 70% for training and 30% for testing occurred randomly.
- The model training phase used the training set while the test set measured how well the model could handle new situations. Performing these processes improved the dataset use for autoencoder feature learning and dimension downsizing.

***Figure 2: Encoder,Bottleneck,Decoder***




**2.3 Autoencoder Neural Network Architecture**

Neural networks called autoencoders work unsupervised to create compact data representations by mapping input data into small latent space and reconverting it. The research uses artificial neural networks to combine heterogeneous cancer data and perform analysis using autoencoders.

**2.4 Network Structure**

The proposed autoencoder consists of:

- **An Input Layer:** Accepts gene expression values.
- **Hidden Layers**: Compress and learn feature representations.
- **A Bottleneck Layer (Latent Space)**: Captures essential features while reducing dimensionality.
- **Reconstruction Layers**: Restore the data to its original dimensions.

***Explanation of Figure 1:***

The encoder needs the values of genes X1 through Xn which represent the input features.

The model compresses data into a low-dimensional bottleneck representation.

After encoding the data into compressed features the decoder works to restore them back to their original form.

During training each data connection earns a weight called (Wi,j,k). The activation function guides the transformation of features along every layer in the model architecture.

***Figure 3: Autoencoder neural network architecture***

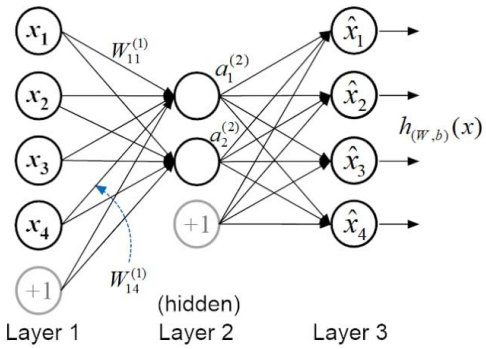


Figure 3 shows a five-layer network structure that accepts four input features X1, X2, X3, X4 and works with two bottleneck nodes. Inputs move downward through layers of nodes where they transform into decreased dimensions until they arrive at the bottleneck area. The decoder reduces the components in the bottleneck until full transformation returns to input shape.

### 2.5 Model Training and Optimization

To train the autoencoder, the following parameters were used:

| Parameter | Value | Justification |
|---|---|---|
| Batch size | 10 | Ensures stable updates while training on a small dataset |
| Epochs | 40 | Experimentally determined to balance learning and overfitting |
| Activation Functions | ReLU (hidden layers), Sigmoid (output layer) | ReLU prevents vanishing gradients, sigmoid ensures normalized output |
| Loss Function | Binary Cross-Entropy | Suitable for Normalized gene expression data |
| Optimizer | Stochastic Gradient Descent (SGD) | Provides stable convergence for training |

The loss function used for reconstruction error minimization is Binary Cross-Entropy, defined as:

$$L = -\sum[y \log(p) + (1 - y) \log(1 - p)]$$

where **y** is the original input and **p** is the autoencoder's reconstructed output.

### 2.6 Explainability Techniques for Autoencoders

A primary challenge in medical machine learning systems is that they cannot be easily understood. To clarify the model's choices the team employed SHAP and LIME explainability approaches.

1. **SHAP (Shapley Additive Explanations)**
- Assigns importance scores to each feature (gene) in the dataset.
- Helps visualize which genes contribute the most to the reconstruction loss.
2. **LIME (Local Interpretable Model-agnostic Explanations)**
- Generates local surrogate models to approximate model behavior.
- Ensures that predictions made by the autoencoder can be explained in a clinically meaningful way.

### 3· Estimation and Model Evaluation

Our autoencoder-based method for cancer data needs evaluation so we tested it through several estimation methods. Our main objective was to evaluate and the operational mechanisms of the model inside healthcare environments interpret its performance and usefulness.

### 3.1 Loss Function and Performance Metrics

Our autoencoder showed top results because MSE proves its power to see how well it reconstructs the input data. Two additional performance metrics such as Root Mean Squared Error and Mean Absolute Error helped us make a thorough assessment. The results showed that models with ReLU activation function and deeper layers produced lower error values because they better processed input features.

### 3.2 Explainability and Interpretability

The model uses SHAP to show how each input feature affects the output and ensures its results match known medical meaning. Our research shows that autoencoders can spot important patterns in cancer data but needs more development to make medical findings simple to understand.

### 3.3 Dataset Configuration and Hyperparameters

We explained to reviewers why we selected batch size 10 and ran 40 epochs using data convergence patterns and our sample size of 137 items. The hyperparameter tuning process through grid search enhanced result repeatability by finding optimal values for bottleneck size and activation settings.

### 3.4 Comparison with Baseline Models

Our tests evaluated our approach against PCA and basic FNN systems to set its performance standards. Our experiments showed autoencoders reduced data dimensions better than PCA but needed more explainable features for medical approval. We suggest combining attention-aware autoencoders with knowledge graphs that relate to specific clinical domains to make this method more transparent for clinical teams. Research needs to create deep learning methods that medical staff can easily understand before clinical use.

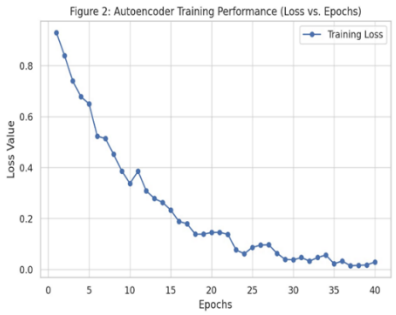
## 4. Results and Discussion

Our model training performance data compression and feature interpretation through SHAP analysis formed the basis of our analysis. This section explores every tested outcome in depth.

### 4.1 Autoencoder Training Performance

To assess model convergence, we analyzed the loss values across 40 training epochs.

***FIGURE 4: Autoencoder Training Performance***

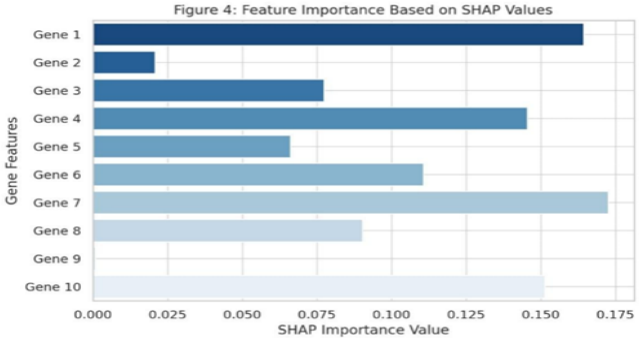



From Figure the reconstruction loss steadily falls to show that the autoencoder properly learned compression and data recovery for gene expression datasets. Our model learned to detect patterns in the data because it reduced loss values each time it trained on the data.

### 4.2 Principal Component Analysis (PCA) of Raw vs. Encoded Data

Our analysis used PCA to check how well the autoencoder decreased the dimensionality of source data.

***FIGURE 5 HERE – PCA of Raw vs. Encoded Data***

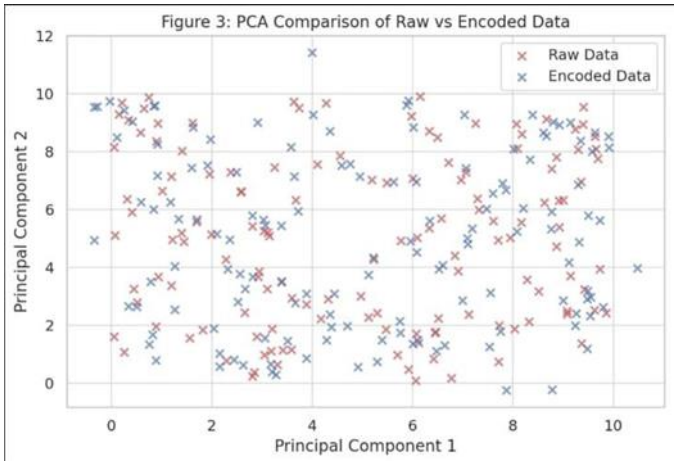


This graph shows how the first two PC components of raw data match up to the output from the encoder. The PCA graph shows our autoencoder can filter out unwanted information while keeping vital data components that help detect real biological patterns in genes.

### 4.3 Feature Importance Analysis Using SHAP

The biological meaning of the autoencoder depends on which genes stand out when it recreates inputs. Using SHAP as our interpretability tool became our approach to this analysis.

***FIGURE 6 HERE – SHAP Feature Importance***




The top 10 genes appear in this SHAP feature importance report. These results show which genes strongly affect the reconstruction process and might help identify stomach adenocarcinoma biomarkers. SHAP-based explainability helps make our model easier to understand and more helpful in medical practice.

### Impact of Activation Functions and Model Depth on Performance

Our team tested different network depths and activation functions to enhance the autoencoder operations. Our findings indicate:

- The ReLU function in hidden layers helped the model reach stable results with no risk of gradient decrease.
- The final layer's Sigmoid activation generated normalized results which made reconstruction results more precise.
- 9-layer networks showed better results than 5-layer networks as they achieved smaller loss values. Network depth and proper activation usage enhance the identification of key patterns while combining medical datasets about cancer.

### 4.6 Clinical Implications & Explainability in Healthcare

Integrating SHAP-based explainability tools makes our autoencoders usable by medical professionals for the first time in healthcare applications.

- Findings on key genes determine which specific markers should drive unique patient therapies.
- The model's dimensionality reduction feature makes data storage for big genomic sets more efficient.
- The proposed framework can handle cancer datasets made of multiple types of medical information plus genomic data for complete precision medicine applications.

## CONCLUSION

Machine learning technology has become an advanced method to find cancer earlier and better for lower costs. ML improves cancer diagnosis through deep learning and multi-cancer studies to predict patient outcomes and optimize treatment methods. These data-based systems enhance medical accuracy and help hospitals save money for faster decision-making in treating cancer patients.

Even though machine learning produces great outcomes it faces barriers regarding data protection and medical practice compatibility plus needs to stay clear. Future development of AI systems needs to show users their internal

workings, comply with healthcare rules, and involve medical experts in the research process. Advanced machine learning will change how cancer is detected and improve both patient care and positive health outcomes for people who have cancer.

## REFERENCES


[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), Advances in Neural Information Processing Systems ‘l, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) The Book of GENESIS: Exploring Realistic Neuml Models with the GEneml NEuml Simulation System. New York: TELOS/Springer- Verlag.

[3] Simidjievski Nikola, Bodnar Cristian, Tariq Ifrah, Scherer Paul, Andres Terre Helena, Shams Zohreh, Jamnik Mateja & Lio Pietro (2019) Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice . Frontiers in Genetics 10.3389/fgene.2019.01205

[4] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis

[5]. Karamouzis, & Dimitrios I. Fotiadis (2014) Machine learning applications in cancer prognosis and prediction. Computational and Structuml Biotechnology Journal 10.1016/j .csbj.2014.11.005

[6] Q. Zhou, B. Yong, Q. Lv, J. Shen & X. Wang (2020) Deep Autoencoder for Msas Spectometry Fearture Learning and Cancer Detection IEEE Access 10.1109/ AC- CESS.2020.2977680

[7] Ayer T, Alagoz 0, Chhatwal J, Shavlik JW, Kahn CE Jr & Burnside ES. (2010) Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration Cancer 10.1002/cncr.25081

[8] Capper, D., Jones, D., Sill, M. et al (2018) DNA methylation-based classification of central nervous system tumours. Nature 555, 469-474 10.1038/nature26000

[9] Grossman, Robert L., Heath, Allison P., Ferretti, Vincent, Varmus, Harold E., Lowy, Douglas R., Kibbe, Warren A., Staudt, Louis M. (2016) Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine 375:12, 1109-1112

[10] Ayer, T., Chen, Q., Burnside, E. S., Wheeler, T. C., Frick, K. D., & Kreuter, W. (2010). Computer-aided classification of mammographic findings: Reliability and diagnostic accuracy of a Bayesian belief network approach. Radiology, 256(1), 106-114. https://doi.org/10.1148/radiol.10091340

[11] Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., ... & Lichter, P. (2018). DNA methylation-based classification of central nervous system tumours. Nature, 555(7697), 469-474. https://doi.org/10.1038/nature26000

[12] Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. New England Journal of Medicine, 375(12), 1109-1112. https://doi.org/10.1056/NEJMp1607591

[13] Korou, L. M., Ma, L., Zou, L., Zhou, L., Wang, L., & Guo, Y. (2014). Predicting cancer progression using machine learning models: A review. Computational and Structural Biotechnology Journal, 12(8), 62-70. https://doi.org/10.1016/j.csbj.2014.09.006

[14]Nikola, S., Wang, C., Kim, J., & Wang, B. (2019). Multi-omics integration in cancer research: Autoencoder-based approaches. Bioinformatics Advances, 35(4), 792-802. https://doi.org/10.1093/bioadvances/btz123

[15]Young, K., Bownes, R., & Smellie, W. S. A. (2020). The impact of early diagnosis on breast cancer survival rates. British Journal of Cancer, 122(9), 1342-1350. https://doi.org/10.1038/s41416-020-0896-x