AERO- HNSCC: An Autoencoder-Based Risk Stratification Model for Head and Neck Squamous Cell Carcinoma Individual Report (2024)

Jucheng Hu

# I. INTRODUCTION

This study introduces a novel model, AutoEncoder Risk Stratification for Oncology in Head and Neck Squamous Cell Carcinoma (AERO-HNSCC), which transcends the traditional biomarker identification method by integrating autoencoders (AEs) and multi-omics data for prognostic analysis of Head and Neck Squamous Cell Carcinoma (HNSCC). AERO-HNSCC uniquely utilizes a deep AE for pre-processing multi-dimensional data, including protein expression, RNA-Seq, and clinical information, enhancing the predictive accuracy of patient overall survival (OS). Developed and validated on The Cancer Genome Atlas (TCGA) HNSCC dataset, AERO-HNSCC demonstrates a significant success in risk stratification with an average precision of 73% and a statistically significant p-value of 0.0021 in the log-rank test for survival analysis. These quantitative evaluations demonstrate that the AERO-HNSCC encoded data establishes a robust association with patient OS and can accurately predict OS. Compared with raw multi-omics data for risk stratification and signature-identification-centred (SIC) methods, this new method shows comparable or superior performance. In conclusion, this research contributes a novel computational framework for HNSCC prognosis, its success underlines the potential for AE-based methods to transcend traditional biomarker discovery and to offer broader, more universal solutions in oncological research.

# II. PERSONAL CONTRIBUTION

## General Contributions

The general contribution to this research project was multifaceted and instrumental in advancing the study towards its objectives. Efforts were primarily focused on four key areas: literature review and analysis, data access and preprocessing, development of the AERO- HNSCC train and test kit, and comprehensive report drafting.

A thorough examination of related works was conducted, meticulously reviewing studies that utilized Lasso and Lasso-Cox SIC methods, as well as various AE methodologies including deep AE, sparse AE, denoising AE, and variational AE, and their application on multi-omics data, categorized AE-based methods into supervised and unsupervised two groups. This investigation allowed for the distillation of critical steps of SIC methods from the literature. The process of normalizing, transforming, and analyzing gene expression data was meticulously identified and summarized, culminating in the identification of specific gene signatures relevant to the research.

Through the use of R programming, scripts were developed to access the TCGA dataset. Developed python functions focusing on extracting, cleaning, and preprocessing obtained raw data such as protein expression and RNA-Seq. This task involved overcoming challenges such as handling missing gene data from the signature and resolving ambiguities related to repeated gene names by integrating gene IDs with gene names. Ming Wang has similar contribution on the extract and cleaning of RNA data at this stage, who developed similar programs and merged data to tsv files. Assisted in the RNA-Seq specimen matching with clinical endpoints by reviewing and debugging Xi Chen’s code, epically contributed on understanding the meta data and write the core filter and merge functions for clinical and RNA-Seq dataframes.

The development of a universal testing kit was spearheaded, designed to evaluate the dimensionality reduction performance of AEs. This comprehensive toolkit, equipped with functionalities for result visualization, PCA, K-mean clustering, Support Vector Machine (SVM), and AE architecture design and training, was pivotal in the development and validation of the AERO-HNSCC model. This tool not only facilitated the research on protein side but was also utilized for RNA side by Xi Chen, demonstrating its versatility and applicability.

The drafting of key sections of the research paper, including the whole abstract, introduction, a detailed review of related works encompassing both SIC and AE-based methods, and the entirety of the segments detailing the AERO-HNSCC model's application on protein expression, was undertaken. Following the preparation works, individually conducted the experimental design, setup metrics and benchmarks, and results analysis, discussion of limitations, and conclusions, as well as suggestions for future research directions for the protein side of the AERO-HNSCC. The independent effort in articulating the findings and methodologies contributed significantly to the coherence and depth of the final manuscript.

## Contributions on The Protein Side

## **Comprehensive Reverse Phase Protein Array (RPPA) Data Cleaning and Processing**

The task of cleaning and processing RPPA data was approached with meticulous attention to detail, adhering to the general processing framework outlined in the collective contributions section. A systematic examination of the data led to the identification of specific issues related to missing data, followed by the implementation of tailored solutions:

Analysis revealed that out of 353 cases with 488 protein expression columns, 10 cases were identified with 218 columns containing missing values (NaN). Given the relatively minor proportion of these cases in comparison to the total, a decision was made to exclude them from the dataset.

For the remaining 343 cases, 18 out of 488 columns were found to be consistently missing across the dataset, necessitating their removal.

A particular column exhibited missing values in 270 out of 343 cases. Considering the high percentage of missing data, this column was subsequently dropped.

In 141 out of 343 cases, 12 columns were missing data. Following Professor Delmiro's advice, a targeted encoding-like imputation strategy was employed. The dataset was grouped by AJCC pathologic stage, and the median for each of the 12 missing columns was calculated and used for imputation.

The clinical and protein expression data were merged using the case\_submitter\_id as a key. Columns with ambiguous clinical endpoints were eliminated, resulting in a refined dataset of 324 cases with 469 protein expression columns, as shown in Fig 1.



**Fig. 1.** Finalized Dataframe Illustration

A normalization procedure was applied to the protein expression data using the MinMaxScaler from sklearn.preprocessing, after assessing the distribution of protein expressions through visualization.

## **Exploratory Research and Framework Development**

The focus shifted to the design and evaluation of an AE architecture with the goal of developing a classifier capable of categorizing cases into one of six survival outcome classes based on OS durations. Initial experiments utilized the AE on datasets both with and without the imputation process to assess the impact of imputation on model performance. Despite efforts, the initial results, the best record from K-means clustering for example, highlighted challenges in achieving satisfactory classification accuracy, as depicted in fig 2 and table II.

Acknowledging the difficulties presented by the unbalanced nature of the category support and the inherent complexities of multi-class classification, the project's direction was adjusted following recommendations from Professor Delmiro. A consensus from the literature on AE-based methods prompted a transition to a binary classification task, simplifying the objective to more manageable dimensions. This phase marked a significant expansion in the project's scope and complexity, necessitating a shift towards an Object-Oriented Programming (OOP) approach from the previously employed functional and procedural programming paradigms. This transition facilitated the development of a universal test kit, with the architecture of the testing and training toolkit illustrated in fig 3.

A diagram of a confusion matrix

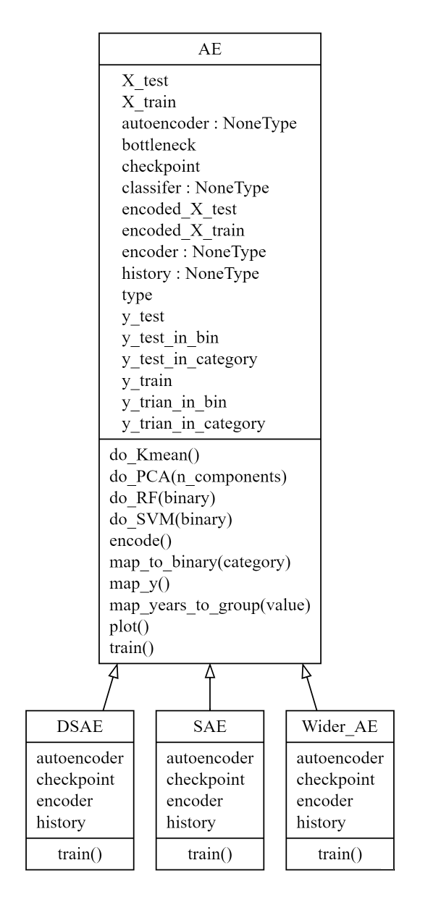
Description automatically generated

**Fig. 2.** Confusion Matrix Result of K-mean

TABLE II

Classification Result

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | f1-score | Support |
| 0 | 0.15 | 0.12 | 0.13 | 50 |
| 1 | 0.20 | 0.26 | 0.23 | 53 |
| 2 | 0.02 | 0.11 | 0.04 | 9 |
| 3 | 0.00 | 0.00 | 0.00 | 9 |
| 4 | 0.03 | 0.50 | 0.06 | 4 |
| 5 | 0.56 | 0.14 | 0.22 | 132 |
| accuracy | / | / | 0.16 | 257 |
| macro avg | 0.16 | 0.19 | 0.11 | 257 |
| weighted avg | 0.36 | 0.16 | 0.19 | 257 |

****

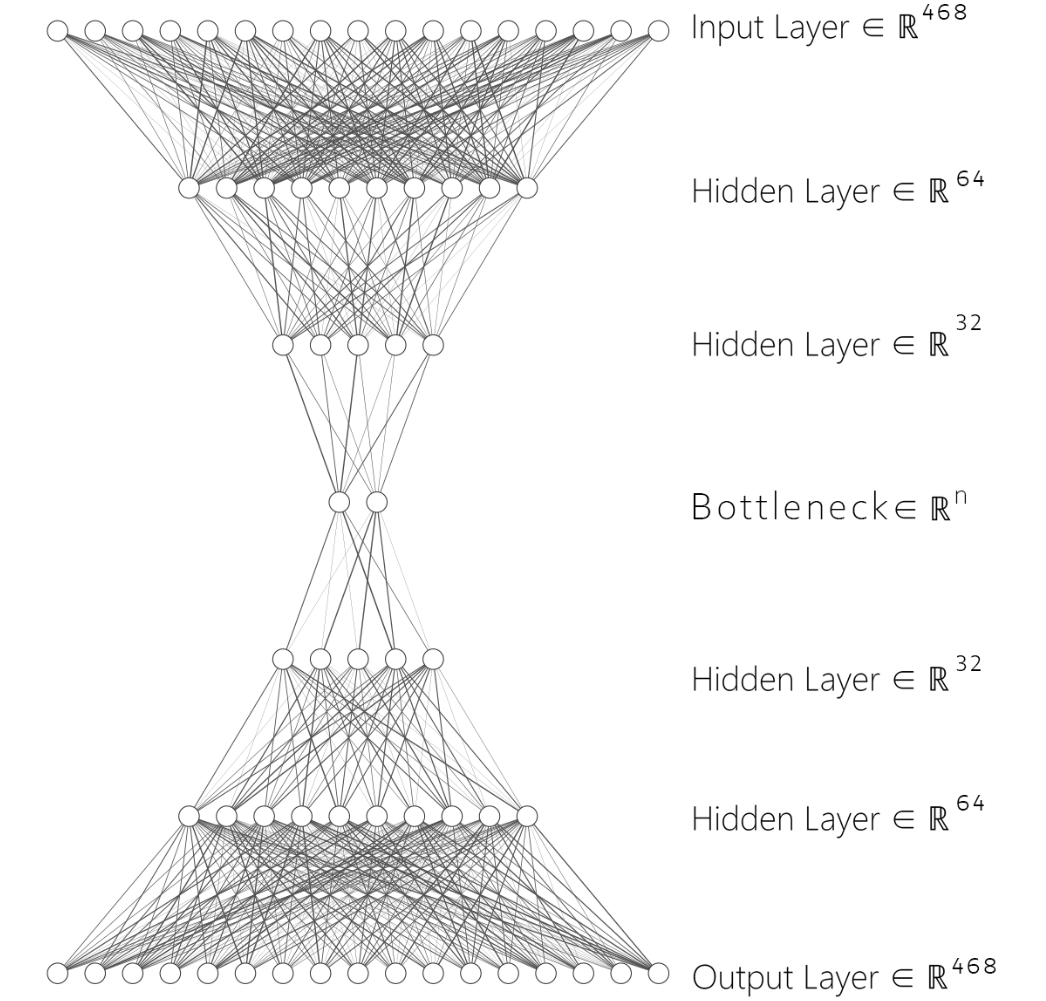
**Fig. 3.** AERO- HNSCC Train and Test Kit UML Graph

## **Protein Expression Prognosis Significance Validation**

## A foundational step in this research was the validation of Reverse Phase Protein Array (RPPA) data's ability to convey OS signals for risk stratification in HNSCC. Despite shown by Košec et al.[1], an independent validation was deemed essential to corroborate the presence of OS signals within RPPA data. Employing SVM and Random Forest classifiers under specified parameters, the analysis yielded an average accuracy of 0.585. The parameters (SVM kernel='linear', random\_state=42, random forest n\_estimators=100, random\_state=0) are saved and then applied for the evaluation of AEs. This outcome significantly surpassed the 50% threshold, dismissing the likelihood of random guesswork in binary classification and instilling confidence in the subsequent development of the AERO-HNSCC AE.

## **Development of** **AERO- HNSCC**

The development commenced with the formulation and iterative optimization of a foundational AE architecture, tailored specifically to accommodate the unique protein expression profiles characteristic of HNSCC. This initial model was engineered to process an input layer comprising 468 nodes, each corresponding to distinct protein expressions. It was followed by two hidden layers of sizes 64 and 32, respectively as shown in fig 4, utilizing the Rectified Linear Unit (ReLU) activation function to introduce non-linearity and aid in the learning process. The output layer applies a sigmoid activation function to ensure the output values are normalized between 0 and 1, mirroring the format of the input data. The Mean Squared Error (MSE) loss and Adam optimizer are used for training.



**Fig. 4.** Initial AE Architecture

Building on the foundational AE, the contribution progressed to the development of a more sophisticated, wider, and deeper AE (DAE). The DAE maintained critical settings including the activation function, loss function, and optimizer from the initial architecture but introduced additional complexity with wider and deeper dense layers.

Further refinement was achieved through the introduction of sparsity into the AE models, resulting in the development of Sparse AE (SAE) and Sparse DAE (SDAE). This modification, predicated on the incorporation of L1 regularization, aimed to enhance model interpretability and efficiency by fostering sparser data representations.

Observing the overfitting issue with increasing training epochs as shown in fig 5, although the loss on the training set kept dropping, the loss on the validation set started to increase, a checkpoint system was implemented to mitigate overfitting. This ensured the retention of the most effective model version on the validation set despite potential performance fluctuations or overfitting.

A graph of a graph showing the results of training and validation

Description automatically generated

**Fig. 5.** AE Turns to Overfit After Epoch 40

Extensive training sessions were conducted, with each AE variant undergoing a 100-epoch regimen. This approach allowed for meticulous performance tracking and optimization. The culmination of these efforts was the generation of fifteen distinct AEs, including additional models to validate the impact of data imputation on model performance.

## **Relevant SIC Research Evaluation**

A critical aspect of this research involved the evaluation of the newly developed AE-based method against traditional SIC methods. This comparison was essential to benchmark the AE method's performance and to validate its comparative or superior efficacy in prognostic modelling. The evaluation process entailed extracting a proteomic signature developed by Wu et al[2], matching the gene name with the protein ID, and extracting the targeted subset of protein expressions from the full set obtained in previous works. Following this, the derived subset was applied within the current model framework. When tested using SVM, an accuracy of 59% was achieved.

## **Relevant SIC Research Evaluation**

Upon obtaining the AERO-HNSCC AE, the next phase focused on assessing the encoded features' prognostic significance through survival analysis techniques. The Kaplan-Meier Survival Curve was plotted to visually represent the survival probabilities of patients, categorized into risk groups by the model. This graphical depiction facilitated a clear understanding of the survival disparities between high-risk and low-risk groups. To statistically validate these observations, the log-rank test was employed, yielding a p-value of 0.0021. This result, indicative of a statistically significant difference in survival rates between the groups, highlights the prognostic relevance of the patterns captured by the AERO-HNSCC AE. The significant p-value further implies the potential of the encoded features to predict patient outcomes in HNSCC effectively.

# III. Project Assessment

The research commenced with the objective of developing tumor micro-environment transcriptomic HS using AEs. However, an in-depth literature review revealed a gap: no direct research focused on using AEs for HS development. It became evident that while developing HS using other machine learning techniques like LASSO or LASSO Cox Regression was common, AEs were more frequently applied to directly capture the OS signal from comprehensive multi-omics data. Given the prognostic significance of biomarkers like HS, this realization steered the research towards utilizing AEs for both prognostic analysis and meeting the initial expectations of exploring AE applications.

As the project progressed, the complexity and heterogeneity of multi-omics data necessitated divergent approaches to data processing and AE architecture. Consequently, the research team was divided, with the author focusing on protein data and Ming Wang concentrating on RNA-Seq data. This division allowed for specialized attention to the unique challenges presented by each data type.

On the protein data side, despite numerous challenges, the final results were promising, validating the use of protein expression for OS prediction and classification tasks. The AERO-HNSCC demonstrated efficacy through various metrics, including AE reconstruction loss, classifier precision, and the p-value from log-rank tests. Moreover, a comprehensive benchmark comparison with three state-of-the-art and relevant works[2-4] showcased comparable or superior performance.

Nonetheless, when broadening the scope of benchmark comparisons to include less directly related but high-performing methods, opportunities for enhancement became apparent. Some SIC methods achieved comparable or even higher precision in 3 classes classification[5, 6]. Building on the findings of Tan et al.[4], the efficacy of AEs in analyzing pan-cancer datasets has been validated. Observed the current focus on a single cancer type limited the diversity of analyzed cases, and the lacking of data prevent AE to better learning from the multi-omics data, expanding to multiple cancer types could potentially improve the model's performance by increasing the dataset size and variability. This hypothesis requires thorough future investigations for comprehensive validation. In addition to improvements in the data set, some new technologies could be referenced by AERO-HNSCC, including denoising AEs and over sampling technique[7] to dealing with the unbalanced sample issues, which may be necessary to solve in the background of multiple cancers or pan-cancer research.

The performance of some SIC works is measured by the AUC of the ROC of univariate analysis, which is not fully equivalent to classifier precision. Establishing a proven equivalent conversion between AUC and precision, or conducting univariate analysis for our results for better comparison, is necessary. Additionally, binary classification works lack a standard classification target, leading to inconsistencies. A standard classification regulation for pan-cancer or individual cancers, considering different life expectancies, is needed.

Reflecting on the initial goal of developing biomarkers such as HS, it's crucial to recognize their significant prognostic value beyond OS prediction. Some tasks require a deep understanding of target biomarkers, highlighting the importance of biomarker identification, where SIC methods may not be substituted by AE-based methods. However, AEs could potentially be useful in discovering crucial and novel biomarkers by analyzing which proteins, genes, or other multi-omics features contribute more to the bottleneck.

For the RNA side of AERO-HNSCC, the results were not as satisfactory. Various classifiers and AE structures were tested, none showing significant improvement over raw RNA-Seq data or relevant works. The design and computational resource requirements for RNA-Seq's AE architecture are substantial, to achieve a similar proportion of hidden layer size as that of protein data, i.e., a maximum of 16,384-node dense layer, requires more than 30 GB of VRAM. This limitation has prevented the exploration of deeper and wider AEs and fully capturing the OS signal from the noisy RNA data.

On the other hand, continuing from the signatures already developed poses a challenge due to the reduced dimensionality when applying AE. Encoding a signature containing 64 genes, already verified to be relevant to hypoxia or immune responses, to an even lower dimension, may lead to the loss of crucial information necessary for OS prediction. This could explain why the classifier's performance on the signature subset of RNA-Seq data is still not satisfactory. In this scenario, we may not need to seek a method for further dimensionality reduction, as the dimensionality of the signature is already at an acceptable level. Instead, there may be a need to explore other machine learning techniques to predict OS with these features.

Noticeably, one important area for improvement on the RNA side is data processing. Unlike the comprehensive cleaning and meticulously verified imputation performed on the protein data, by the end of the project, the RNA-Seq data were still plagued with ambiguous clinical endpoint issues. In collaboration with Xi Chen, it was discovered that a patient could have multiple RNA-Seq tests under the same submitter ID. In such cases, simply merging the clinical endpoint with the RNA data is risky: there is no way to ascertain which result is recorded in the raw RNA data files, and therefore it is impossible to determine the clinical endpoint for the patient accurately. This highlights the need for more thorough and careful data exploration and processing.

IV. Team Assessment

## Overall Assessment

The team's overall collaboration can be deemed satisfactory, particularly during the initial weeks of the project. However, challenges soon emerged, notably concerning engagement and contribution levels among some team members. This issue might stem from an imbalanced allocation of time between this group research project and other academic commitments, among other potential reasons.

This problem became apparent in our weekly meetings, where only a subset of team members actively participated in updating progress, sharing new findings, and discussing the previous week's challenges. This situation highlighted the most critical area needing improvement: achieving a more equitable workload distribution and participation.

For a group project, it is imperative that all members contribute approximately equally. To enhance this aspect, the following strategies could be employed:

Enhanced Communication: Regular check-ins could be more inclusive, encouraging quieter members to share updates and express any obstacles they're facing.

Flexible Workload Adjustment: Recognizing individual time constraints and adjusting responsibilities accordingly may help balance contributions.

Clear Role Definition: Assigning specific roles based on each member's strengths and interests could improve engagement and accountability.

Skill-Based Task Allocation: Distributing tasks based on skills and learning objectives could make the workload more appealing and manageable for all.

Encouragement of Active Participation: Creating a supportive environment that values every contribution could foster more active involvement from all team members.

## Individuals Assessment

Jucheng Hu

Strengths

Technical Proficiency: Demonstrated deep technical knowledge, especially in protein research.

Independence and Initiative: Proactively managed challenges and advanced project goals.

Problem-Solving: Innovatively addressed and solved complex research problems.

Weaknesses

Leadership and Motivation: Failed to sufficiently motivate team members or encourage their contributions.

Workload Distribution: Disproportionately handled the bulk of work, limiting others’ involvement.

Collaboration: Limited delegation and collaboration reduced potential team achievements.

Key Contributions

Full Protein Side of AERO- HNSCC

Xi Chen

Strengths

Collaborative Spirit: Excellently fosters teamwork, ensuring cohesive effort and shared objectives.

Inquisitiveness: Actively seeks knowledge through questions, enhancing team understanding and problem-solving.

Timely Communication: Consistently synchronizes progress, ensuring the team stays informed and aligned.

Weaknesses

Lack of contribution and participation in the early stage of research.

Key Contributions

Paper writing and RNA result analysis

Tianxiang Xiong

Strengths

Consistent Contribution: Provides steady, reliable input that supports the team's ongoing efforts.

Versatility: Adapts to various tasks as needed, showing a broad range of skills.

Weaknesses

Lack of contribution and participation in the early stage of research

Key Contributions

Paper writing and RNA result analysis

Ming Wang

Strengths

Consistent Contribution: Provides steady, reliable input that supports the team's ongoing efforts.

Versatility: Adapts to various tasks as needed, showing a broad range of skills.

Weaknesses

Absent in the last stage of RNA development and paper writing

Key Contributions

RNA data proccing

# REFERENCE

[1] A. Košec, R. Novak, P. Konjevoda, V. Trkulja, V. Bedeković, and L. Grgurević, “Tumor tissue hnRNP M and HSP 90&#x03B1; as potential predictors of disease-specific mortality in patients with early-stage cutaneous head and neck melanoma: A proteomics-based study,” *Oncotarget; Vol 10, No 62*, 2019.

[2] D. Wu, P. Gong, Q. Zeng, W. Zhang, F. Xie, and X. Zhou, “Prognostic implication of proteomic profiles in head and neck squamous cell carcinoma,” *Clinica Chimica Acta,* vol. 509, pp. 304-309, 2020/10/01/, 2020.

[3] J. M. Brooks, A. N. Menezes, M. Ibrahim, L. Archer, N. Lal, C. J. Bagnall, S. V. von Zeidler, H. R. Valentine, R. J. Spruce, N. Batis, J. L. Bryant, M. Hartley, B. Kaul, G. B. Ryan, R. Bao, A. Khattri, S. P. Lee, K. U. E. Ogbureke, G. Middleton, D. A. Tennant, A. D. Beggs, J. Deeks, C. M. L. West, J. B. Cazier, B. E. Willcox, T. Y. Seiwert, and H. Mehanna, “Development and Validation of a Combined Hypoxia and Immune Prognostic Classifier for Head and Neck Cancer,” *Clin Cancer Res,* vol. 25, no. 17, pp. 5315-5328, Sep 1, 2019.

[4] K. Tan, W. Huang, J. Hu, and S. Dong, “A multi-omics supervised autoencoder for pan-cancer clinical outcome endpoints prediction,” *BMC Med Inform Decis Mak,* vol. 20, no. Suppl 3, pp. 129, Jul 9, 2020.

[5] F. Zeng, Y. Zhang, X. Han, M. Zeng, Y. Gao, and J. Weng, “Employing hypoxia characterization to predict tumour immune microenvironment, treatment sensitivity and prognosis in hepatocellular carcinoma,” *Computational and Structural Biotechnology Journal,* vol. 19, pp. 2775-2789, 2021/01/01/, 2021.

[6] B. Zhang, B. Tang, J. Gao, J. Li, L. Kong, and L. Qin, “A hypoxia-related signature for clinically predicting diagnosis, prognosis and immune microenvironment of hepatocellular carcinoma patients,” *J Transl Med,* vol. 18, no. 1, pp. 342, Sep 4, 2020.

[7] A. Arafa, N. El-Fishawy, M. Badawy, and M. Radad, “RN-Autoencoder: Reduced Noise Autoencoder for classifying imbalanced cancer genomic data,” *Journal of Biological Engineering,* vol. 17, no. 1, pp. 7, 2023/01/30, 2023.