Journal of Biosciences and Medicines, 2025, 13(3), 155-182
https://www.scirp.org/journal/jbm
ISSN Online: 2327-509X
ISSN Print: 2327-5081

Gene Expression Model for the Disease Prediction with Auto-Encoder Model with Classifiers

Arjun Kunwar, Shulin Wang

School of Computer Science and Technology, Hunan University, Changsha, China
Email: arjunkunwar5@gmail.com, developer.alok3@gmail.com

How to cite this paper: Kunwar, A. and Wang, S.L. (2025) Gene Expression Model for the Disease Prediction with Auto-Encoder Model with Classifiers. *Journal of Biosciences and Medicines*, **13**, 155-182.
https://doi.org/10.4236/jbm.2025.133013

Received: December 23, 2024
Accepted: March 11, 2025
Published: March 14, 2025

Abstract

Gene expression is the process through which genetic information in DNA is converted into functional products, primarily proteins. This involves two main steps: transcription, where DNA is copied into messenger RNA (mRNA), and translation, where mRNA is decoded by ribosomes to synthesize proteins. Gene expression is tightly regulated to ensure proper cellular function, and its analysis is vital in fields like cancer research, drug development, and genetic engineering. Hence, this paper proposed effective Voting-based Stacked Denoising Auto-encoder (VSDA) for the prediction of diseases. The VADA model uses the stacked model within the Auto-encoder for the accurate prediction of the gene expressions. This paper investigates the performance of four machine learning classifiers—Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbours (KNN), and Multi-Layer Perceptron (MLP)—on a cancer diagnosis dataset, using metrics such as Precision, Recall, F1-Score, and Support across multiple cancer types. Our results show that MLP achieves the highest overall performance with an average Precision of 0.92, Recall of 0.75, and F1-Score of 0.74. SVM follows closely with an average Precision of 0.89, Recall of 0.78, and F1-Score of 0.79, demonstrating strong reliability, particularly for cancers such as LUAD, KIRC, and THCA. RF exhibited an average Precision of 0.75, Recall of 0.68, and F1-Score of 0.66, indicating balanced performance but with slightly lower accuracy compared to SVM and MLP. KNN, while performing well in certain cancer types, had the lowest overall F1-Score of 0.60 and Precision of 0.71, showing greater variability across different cancer types. These results underscore the superiority of MLP in most scenarios, with SVM offering a competitive alternative for specific cancers. The study highlights the importance of classifier selection based on specific cancer datasets, with the goal of improving diagnostic accuracy and supporting clinical decision-making.

Keywords

Auto-Encoder, Stacked Voting, Classification, Cancer Diagnosis, Gene Expression, Prediction

1. Introduction

Gene expression is defined as the use of gene information with a view of synthesizing a functional gene product, which may be a protein or RNA molecule [1]. This process involves two key stages: Transcription and translation. In transcription DNA sequence of a gene is transcribed into messenger RNA (mRNA) in the nucleus [2]-[5]. In translation, the mRNA is than used as a template to construct a corresponding protein in the cytoplasm. Transcription factors, environmental signals, and epigenetic changes act as the factors controlling gene expression intensity and time [6]. Most of the genes implicated in the prediction and diagnosis of cancer exhibit changes in the process of the development of carcinoma. The different genes in tumor cells may be expressed in different levels to that of normal cells with some genes being expressed higher than normal while others lower [7]. They can cause uncontrolled cell growth, the ability to escape through apoptosis and the ability to form metastasis.

Examining gene expression data from cancer samples allows investigators to find objective indicators for cancer existence, its subtype, and its stage. Such enhancements in tools and technologies such as microarrays and RNA-seq facilitate large-scale measures of molecular activity to gain an understanding of the molecular basis of cancer [8].

Auto-encoders, a type of artificial neural network, are increasingly used in gene expression analysis in disease prediction especially in biomarkers detection and disease prognosis [9]. In this context, auto-encoders are types of unsupervised learning techniques that can reduce data, especially gene expression profiles contained in this study, into a smaller form then attempt to reconstruct the data as close as possible to the original data [10]. The encoder side of the auto-encoder codes the actual gene expression data into a format that saves only the most important aspects, patterns or features of the given data set and the decoding side of the auto-encoder tries to reconstruct the input data [11]. If required for disease prediction the auto-encoders can then trained on the gene expression data from both the healthy and diseased samples giving the model the ability to learn the features of gene expression which are likely a sign of the disease type of interest such as cancer, diabetes, or neurological disorders [12]. The learned representations can then be used for activities such as novelty detection, classification or clustering, aimed at the identification of new biomarkers, prognosis of disease and diagnosis. Auto encoders is one of the deep learning models that is getting popular for disease prediction in genomic data specially gene expression data [13]. These models are intended to effectively capture a low-dimensional representation of

high-dimensional data which may include gene expression profiles by encoding the input data and then decoding it. The encoder and decoder take accounts for definite values of gene expression in healthy and diseased states with respect to gene expression data. Auto-encoders are helpful in reaching the goals of identifying the differences in gene expression related to diseases, discovering biomarkers, and diseases classification by paying attention to the most significant patterns [14]-[16]. Auto-encoders have made significant contributions to cancer prediction by offering an efficient way to analyze complex, high-dimensional biological data such as gene expression profiles, medical images, and proteomic datasets. As unsupervised learning models, auto-encoders reduce dimensionality by learning compact, latent representations of input data while retaining its most informative features. This enables the identification of subtle patterns or anomalies that may indicate the presence of cancer. By leveraging these latent representations, auto-encoders can improve the accuracy of downstream classification tasks, such as differentiating between cancerous and non-cancerous samples or identifying specific cancer subtypes. Additionally, their ability to denoise data and handle missing values enhances the quality of predictions, making them particularly valuable in clinical settings where data variability is common. The methods are especially useful in simplifying complex gene expression data, increasing readability, and identifying novel patterns such as signs of an emerging disease. The compressed representation learned by the auto-encoder can be then used for predictive tasks like patient's class or disease prognosis [17]. Additionally, auto-encoders enable the integration of multi-omics data, providing a more comprehensive understanding of diseases. Despite challenges in model interpretability and generalization across different datasets, auto-encoders hold significant potential in advancing personalized medicine, predicting treatment responses, and discovering novel therapeutic targets [18].

The gene expression datasets for cancer prediction include comprehensive activities of genes in cancer and non-cancer tissues, allowing researchers to establish molecular patterns of the associated diseases [19]. These datasets commonly contain the quantitative values of thousands of genes in the hope of detecting differential patterns between healthy and cancer cells. In cancer prediction, these gene expression datasets mention organizations above are used in combination with the machine learning technology to identify the trivial variations in gene activity that could predictive cancer, its type, or phase [20]. For instance, activation or down regulation of any specific genes can be potential sign of oncogenes like growth independence, anti-apoptosis and cell migration. In these studies, it is possible to discover novel diagnostic markers for primary cancer, prognosis of the disease, and the therapeutical outcome. Popular cancer gene expression datasets such as TCGA bring valuable data into the process of building accurate prognostic models as well as improving a patient's treatment plan by correlating gene expression patterns of tumours to specific clinical outcomes. These datasets are critical in the enhancements of the knowledge we have or the cancer biology besides en-

hancing the reliability of the diagnostic tools that can at long last enhance the quality of services rendered to patients.

This paper proposed voting-based stacked Denoising Auto-encoder (VSDA) model for the prediction and classification of the gene expression. The proposed VSDA model uses the denoising Auto-encoder for the gene expression prediction and classification. This paper investigates the performance of four machine learning classifiers—Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbours (KNN), and Multi-Layer Perceptron (MLP)—on a cancer diagnosis dataset, using metrics such as Precision, Recall, F1-Score, and Support across multiple cancer types. Our results show that MLP achieves the highest overall performance with an average Precision of 0.92, Recall of 0.75, and F1-Score of 0.74. SVM follows closely with an average Precision of 0.89, Recall of 0.78, and F1-Score of 0.79, demonstrating strong reliability, particularly for cancers such as LUAD, KIRC, and THCA. RF exhibited an average Precision of 0.75, Recall of 0.68, and F1-Score of 0.66, indicating balanced performance but with slightly lower accuracy compared to SVM and MLP. KNN, while performing well in certain cancer types, had the lowest overall F1-Score of 0.60 and Precision of 0.71, showing greater variability across different cancer types. These results underscore the superiority of MLP in most scenarios, with SVM offering a competitive alternative for specific cancers. The study highlights the importance of classifier selection based on specific cancer datasets, to improve diagnostic accuracy and support clinical decision-making.

2. Related Works

There has been much interest in the last ten years in research on gene expression and cancer risk assessment since genomic technologies can offer the identification of molecular pathways and disease markers. Tremendous studies have been produced to use molecular biology prospectives and different machine learning and statistical methods to predict the gene expression for capturing disease, progress, and responses to therapy. Examples of work in this area are discussed below: classifiers such as support vector machines (SVM), random forest, and deep learning algorithms, dimensionality reduction; principles such as principal component analysis (PCA) auto-encoder. Such strategies have been employed for discovering biomarkers that may be useful for early detection of disease and individualised therapies.

Despite the several promising uses of auto-encoders and deep learning models for gene expression analysis in disease prediction, there are some challenges that should be responded to however, one of the main disadvantages of the described methods is that they are not easy to interpret. Although auto-encoders are good at dimensionality and capturing high level representations of gene expression data they induce non-interpretable representations which can make it hard to associate them to particular biological processes or pathways. There are also reporting differences that make difficult their acceptance in environments such as the clinic, where the biological relevance of these features, predictive for the disease, must be

clear. The last drawback is data variation; gene expression data is highly variable based on the platform used, the cohort sampled, and the experimental setting looked into, which in turn influences the scalability of the deep learning results. This approach is essential given that models trained on one dataset can have poor performance when applied to another thus the call to address data heterogeneity. Another issue is overfitting, which is vital when using large-dimensional features like gene expressions; a model falls in love with the noise. For this, it is required to use the regularization techniques and proper cross validation which, however, consumes a lot of CPU time. Further, the requirement for big volumes of annotated data and creating a robust pipeline of deep learning models continue to persist, as it might be highly costly and time-consuming to gather high-quality labelled datasets for rare disease or specific type of cancer. Last, there is need to work on Deep learning model such as auto-encoders that still pose great computational demands as means of data storage hence may not be readily applicable in small research groups or even in resource confined settings.

3. Proposed Voting-Based Stacked Denoising Auto-Encoders (VSDA)

The proposed model was studied for VSDA using an autoencoder that works based on voting to self-predict cancer. During this phase, patterns were identified and predictions were evaluated using different classifiers. Initially, this research uses ribonucleic acid (RNA) sequencing to identify and analyze changes in gene expression patterns (transcriptome) within cancerous cells. Those are pre-processed and denoised for further processing. Within the pre-processed data to reduce the dimensionality of data principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE). With the Voting-based Stacked denoising auto-encoder model clustering of data sequences is performed to estimate feature extraction and selection. With the estimated features voting-based model is implemented for the prediction of the cancer genes. Once the classification is performed it is evaluated with the different classifiers such as SVM, RF, KNN and MLP. With the classification is performed clustering is performed for the estimation of the attributes in the data. The proposed VSDA model for the prediction is presented in Figure 1.

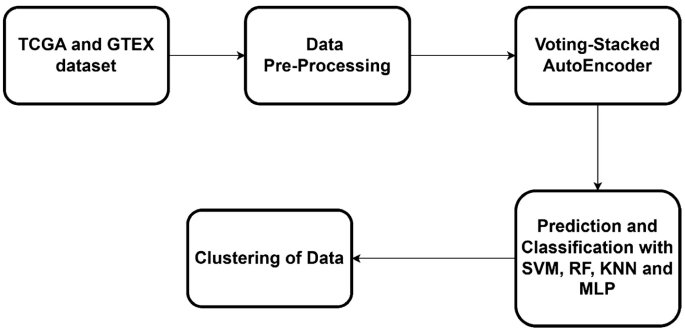


Figure 1. Flow of proposed VSDA.

The proposed flow in Figure 1: presented the flowchart illustrates a structured machine learning pipeline designed to analyze the TCGA (The Cancer Genome Atlas) and GTEX (Genotype-Tissue Expression) datasets. The process begins with the acquisition of these datasets, which are renowned for their extensive genomic and tissue-specific data, serving as a foundation for biological and medical insights. The first step involves data pre-processing, where raw data is cleaned, normalized, and prepared for further analysis. This step ensures the removal of noise and inconsistencies, enhancing the quality and reliability of the dataset. Once pre-processed, the data undergoes feature extraction through a voting-based stacked auto-encoder. This advanced method leverages deep learning to capture complex, hierarchical features, combining the strengths of stacked auto-encoders and ensemble techniques to improve robustness and accuracy. Following feature extraction, the data is clustered to group similar data points, enabling the identification of hidden patterns and relationships. Finally, the clustered data is fed into a prediction and classification module, which utilizes a diverse set of machine learning algorithms—Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP). These models work collaboratively to classify data and predict outcomes with high accuracy, leveraging their individual strengths. This comprehensive workflow integrates data pre-processing, feature extraction, clustering, and classification to provide meaningful insights, particularly in the context of genomic and tissue-specific research.

The Voting-based Stacked Denoising Auto-encoders (VSDA) model offers a unique take on auto-encoders, dimensionality reduction, ensemble learning while providing potential to automate cancer prediction by gene expression analysis. This model is intended to be used in RNA sequencing where profitaciouse details of the cancer cells' transcriptome are delivered. The first step in the proposed approach is Data cleaning; RNAss data pre-processing step which excludes noise and other unimportant characteristics by applying Denoising Auto-encoders (DAE). As these denoising auto-encoders are trained for reconstruction of data with less noise, these reconstructed gene expression patterns serve better for the next step analysis. The final step in the current model entails the process of dimensionality reduction entail Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) to maximize usability of the various data sets and enhance feature extraction. Taking this into account PCA can be used for identifying principal components which describe maximum variance in the data while t-SNE is great for visualizing the data with low dimensions while keeping the neighbors' relationships. These techniques make it easier to eliminate dimensionality and retain only those Gene features that must be analyzed.

The main of the proposed method is based on the Voting-based Stacked Denoising Auto-encoder (VSDA) shown in Figure 2. The trained SDA is a feed forward neural network where every layer is trained to extract rather abstract features of the gene expression data. Each auto-encoder is designed to be trained to mimic input-output mapping with a high degree of accuracy, and hence retain the de-

sired gene patterns in the process. The stacked structure enables the model to develop multiple levels of representation of the data, where the first levels identify straightforward gene expression patterns, while the other levels learn more complex patterns associated with cancer development. Once the features are extracted, the Voting-based approach is then used to ensembles the results of multiple models or classifiers from a list of models/classifiers, such as k-nearest neighbor, naive bayes, support vector machine, and multi-layer perceptron. Ensemble learning technique of generating a final prediction by combining multiple models including SVM, RF, KNN and MLP. As for a voting theory, each classifier will give out a call on the predicted class while the class with majority votes will be passed out as the final decision. This method assists in avoiding individual shortcoming of each classifier and therefore makes the prediction model more accurate.

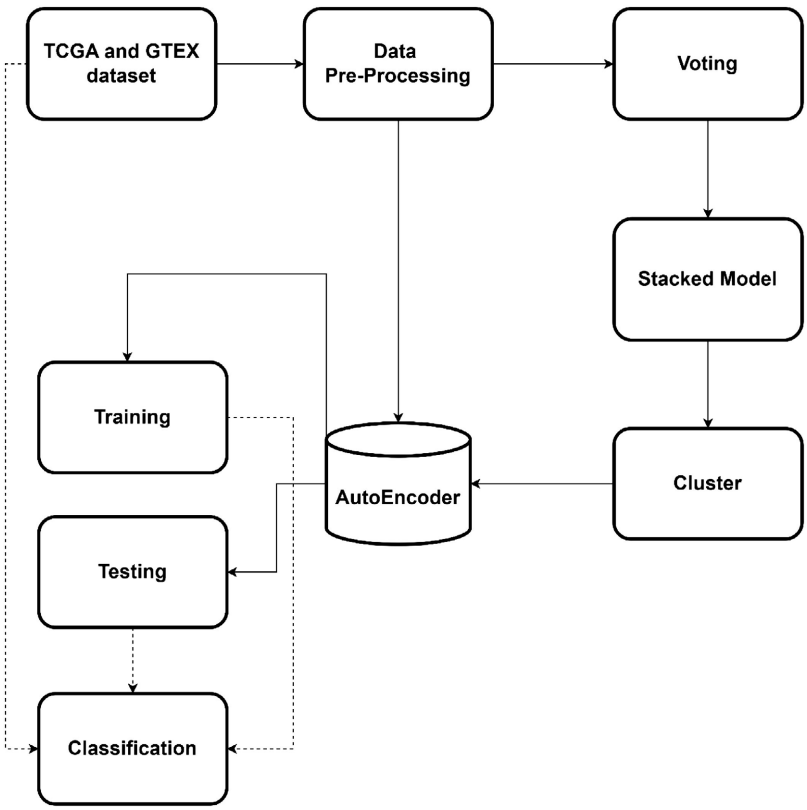


Figure 2. Architecture of proposed VSDA.

3.1. Steps in Voting-Based Stacked Denoising Auto-Encoders

The Voting-based Stacked Denoising Auto-encoders (VSDA) model can be described as a multi-step process to utilize denoising auto-encoders, dimensionality reduction, feature extraction, and classification through voting technique. The first process involves filtering the features and cleaning the RNA sequencing data to minimise noise. The above is made possible by the use of Denoising Auto-encoders (DAE). A denoising auto encoder tries to map a noisy input data towards the clean corresponding input data for preventing noise data and making the data

cleaner. For a given input x, a noisy version $\hat{x}$ is generated, and the goal of the auto-encoder is to minimize the reconstruction error: A type is the VSDA model denoising auto-encoder to reconstruct an input vector x from a noisy version $\hat{x}$ where the model parameters are chosen as follows With this, the loss function is defined as in Equation (1)

$$L_{DAE}\left(x_i, \hat{x}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2 \tag{1}$$

In Equation (1), $\hat{x}$ is the reconstruction of x, and x_i and $\hat{x}_i$ are the individual elements of the input and output vectors, respectively. Where $\hat{x}$ is the reconstructed output, and $\hat{x}_i$ are the individual elements of the noisy input and the reconstructed output, respectively. The encoder learns a mapping $g_{\varnothing}$ from the noisy input to a hidden representation h, and the decoder reconstructs the original data represented in Equation (2)

$$\hat{x} = g_{\varnothing}\left(h\right) \tag{2}$$

In Equation (2), $g_{\varnothing}$ are the encoder and decoder functions parameterized by θ and ϕ, respectively. After denoising, dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are applied to reduce the high-dimensional gene expression data into a lower-dimensional space while preserving the most important features. PCA reduces the dimensionality of the data by finding the principal components W that maximize the variance of the data. The projection of the data X on these components stated in Equation (3)

$$X' = X \cdot W \tag{3}$$

where X' is the transformed data, and W is the matrix of principal components. t-SNE aims to find a low-dimensional representation of the data Y that preserves local similarities. After the feature extraction process, the next step involves classification. The features extracted from the stacked denoising auto-encoders are passed to a set of classifiers, including Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP). These classifiers predict the cancer class based on the learned features. It minimizes the following cost function stated in Equation (4)

$$C\left(Y\right) = \sum_{i \neq j}\left\|P_{ij} - Q_{ij}\right\|^2 \tag{4}$$

In Equation (4), P_{ij} represents the probability that a pair of points $\left(i, j\right)$ are neighbors in the high-dimensional space, and Q_{ij} is the corresponding probability in the low-dimensional space. For classification, let the predictions from individual classifiers be $y_1, y_2, \cdots, y_k$. The final prediction $\hat{y}$ is determined by the majority vote using Equation (5)

$$\hat{y} = argmax_c \sum_{i=1}^{k} I\left(y_i = c\right) \tag{5}$$

In Equation (5), $I\left(y_i = c\right)$ is an indicator function that equals 1 if y_i is the class c, and 0 otherwise. The last prediction is the class which is voted for with

the highest frequency. After delivering the cancer classification, the performance of the classifier is measured using various well known classification parameters like accuracy, precision, recall and F1-score. Due to the fact that the classifier plays a central role in the VSDA model, its performance is evaluated comparatively using benchmark models such as SVM, RF, KNN, and MLP. The next step is clustering in which the extent of the attributes of data is estimated. This involves clustering, which is a technique of spreading similar data through the features and the classification accomplished. During clustering the key relation in the data that are not visible can be identified and hence the model can be modified for predicting more accurately. The clustering can be represented by minimizing a clustering cost function, stated in Equation (6)

$$L_{cluster}\left(X,\hat{X}\right)=\sum_{i=1}^{n}\sum_{j=1}^{k}\left\|X_i-\mu_j\right\|^2 \tag{6}$$

In Equation (6), μ_j denotes the centroid of the j-th cluster, and X_i corresponds to the data points in each cluster. The proposed Voting-based Stacked Denoising Auto-encoders (VSDA) works by stacking auto-encoder layers for features, employing dimensionality reduction techniques for data input, and cancer classification done through voting from ensembled layers. Thus, the integration of these approaches within the VSDA model should enhance the quality of cancer prediction driven by RNA sequencing data. By stacking numerous auto-encoders, the paper was able to identify abstract features in data, while through the voting mechanism, the final prediction was reached by most classifiers, and hence minimizing overfitting.

3.2. Data Set

The dataset utilized for the proposed VSDA model are presented as follows:

3.2.1. GTEX and TCGA Data

The Data used in this study was sourced from The Cancer Genome Atlas. This is the largest global repository of genomic data related to cancer. TCGA was initiated in 2006 as a collaborative project spearheaded by the National Cancer Institute's Center for Cancer Genomics and the National Human Genome Research Institute. This vast set of data has furthered many other bodies of cancer research in areas of accurate diagnosis, effective treatment options, and preventive measures. The public accessibility of TCGA data has also significantly helped numerous scientists in their research endeavors, by having ready access to high-quality genomic information for inclusion in their research studies.

The TCGA dataset comprising 8293 samples that were classified by one of the 15 unique types of cancers. This rich dataset was consequently applied to train a machine learning model that was then validated against an independent dataset formed by using Genotype-Tissue Expression project. The GTEX project is an initiative toward forming a public resource that will allow the exploration of tissue-specific expression and regulation of genes for various tissues in the human body.

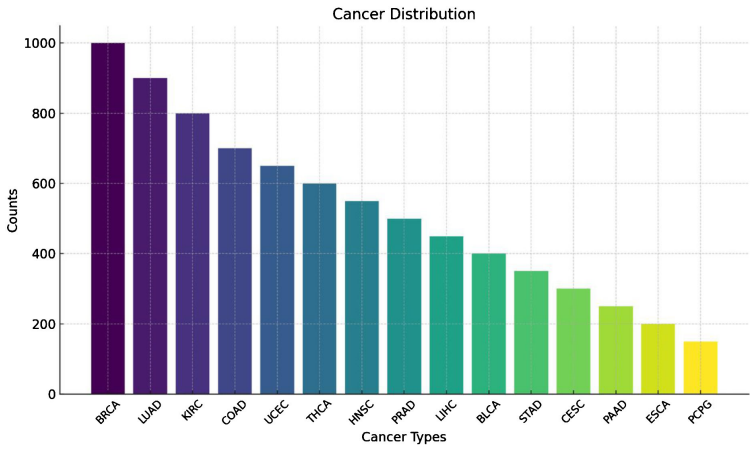
3.2.2. Description of Data

The proposed VSDA model comprises are 38,019 features along with 8293 samples acquired from the TCGA dataset. It is presented in .csv for easier use with machine-learning algorithms. The first column of this dataset is allocated for patient IDs, and the target variable—"Type" is the type of cancer that one wants to predict. The remaining columns would then describe the gene expression values for a variety of genes that would form the basis of the input features that one uses to classify the samples. The target variable describes what kind of cancer it is, where each of the 15 cancer types has been assigned a particular label. A description of the labels and the cancer types themselves can then be described based on the notation used for TCGA, which appears in Table 1.

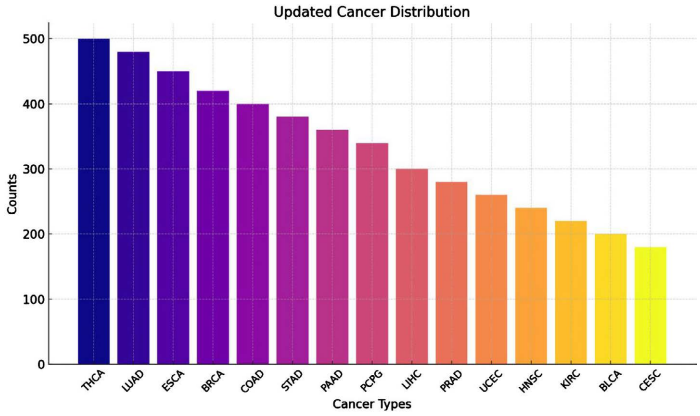
In addition, Figure 3(a) and Figure 3(b) illustrate the spread of cancer types in the TCGA data set with information about the incidence of each kind of cancer within the samples. Figure 3 depicts the spread of GTEX data showing variability in expression among the tissues.

Table 1. Attributes of dataset.

Code	Cancer Type
BRCA	Invasive breast carcinoma
LUAD	Adenocarcinoma of the lung
KIRC	Renal papillary cell carcinoma (kidney)
COAD	Adenocarcinoma of the colon
UCEC	Endometrial carcinoma of the uterine corpus
THCA	Carcinoma of the thyroid
HNSC	Squamous cell carcinoma of the head and neck
PRAD	Carcinoma of the prostate
LIHC	Hepatocellular carcinoma (liver)
BLCA	Urothelial carcinoma of the bladder
STAD	Adenocarcinoma of the stomach
CESC	Squamous cell carcinoma of the cervix and endocervical adenocarcinoma
PAAD	Adenocarcinoma of the pancreas
ESCA	Carcinoma of the esophagus
PCPG	Pheochromocytoma and paraganglioma



(a)



(b)

Figure 3. VSDA analysis with dataset (a) TCGA; (b) GTEX.

This all-inclusive description of the TCGA dataset makes pretty clear how important it is for knowledge in the domain of cancer genomics, based on the emphasis provided on being a foundation to build models of predictions aimed at improvements in cancer diagnosis and treatment strategies.

4. Data Processing with VSDA

The data processing phase in the Voting-based Stacked Denoising Auto-encoders (VSDA) model plays a pivotal role in transforming raw gene expression data into meaningful features that can be used for effective cancer prediction. This step involves a series of pre-processing, dimensionality reduction, feature extraction, and feature selection techniques to ensure that the data is ready for classification.

4.1. Data Pre-Processing

The first step in data processing is to collect and preprocess the raw RNA sequencing data. The data typically consists of gene expression levels across different genes for various samples (healthy and cancerous). Raw RNA sequencing data often contains noise, missing values, and irrelevant features, which need to be cleaned to ensure accurate predictions. This step adjusts the data to ensure that the measurements across different samples are comparable. Common normalization techniques include log-transformation, quantile normalization, or z-score normalization. For normalization, if the expression value for gene g in sample s is denoted as x_{gs}, then the z-score normalization can be defined as in Equation (7)

$$x_{gs}^{normalized} = \frac{x_{gs} - \mu_g}{\sigma_g} \tag{7}$$

In Equation (7), μ_g and σ_g are the mean and standard deviation of the expression levels of gene g across all samples.

4.2. Feature Selection

Feature selection is one of the critical operations in the machine learning pipeline,

focusing particularly on reducing the number of input variables: relevant variables are kept, and irrelevant variables discarded within the feature set, supposedly possessing a strong relationship to the target output variable. The proposed VSDA model incorporates clustering-based feature selection is a method that identifies and selects the most relevant features by grouping similar features into clusters based on their characteristics or relationships with the target variable. In this approach, features within each cluster are highly correlated or exhibit similar patterns, while clusters themselves are distinct. From each cluster, representative features are selected to minimize redundancy and retain critical information. The core purpose of feature selection is the improvement of model performance by selecting meaningful features that give significant contributions to the prediction of the target labels. Feature selection can reduce model complexity and decrease overfitting effect toward improving Acc and interpretability of a model by the elimination of irrelevant or redundant features. In this thesis, the SelectKBest algorithm was applied from the scikit-learn (sklearn) library. This algorithm keeps the k best features by a scoring function. The chi-squared function was applied in our implementation: stating the statistical relationship of all possible pairs of features. Precisely, the method measures how well each feature is correlated with the target variable: it calculates the chi-squared statistic. We also used the F-score for finding the best feature, combining the results of both scoring metrics to achieve robustness. Through the process of feature selection, the original feature set was reduced down to 832 selected features, which were then used in training and evaluating different machine learning models. A sparsity constraint can be added by modifying the loss function to include a penalty term that encourages sparsity in the hidden representation h. For instance, the sparsity constraint can be formulated as in Equation (8)

$$L_{sparsity} = \lambda \sum_{j=1}^{m} \left| \frac{1}{n} \sum_{i=1}^{n} h_{ij} - \rho \right| \tag{8}$$

In Equation (8), h_{ij} is the activation of the *j*-th neuron in the *i*-th sample, mmm is the number of neurons in the hidden layer, and ρ is the desired average activation (usually close to 0). The regularization parameter λ controls the strength of the sparsity constraint. This encourages the auto-encoder to learn a compact and efficient representation by activating only a few neurons, which indirectly leads to selecting important features. Another approach to feature selection is to use statistical measures such as mutual information or correlation to rank the features based on their relevance to the output variable (in this case, cancer classification). For each feature x_i in the dataset, the mutual information $I(x_i, y)$ between the feature and the class label y can be computed using Equation (9)

$$I(x_i, y) = H(x_i) + H(y) - H(x_i, y) \tag{9}$$

In Equation (9), $H(x_i)$ and $H(y)$ are the marginal entropies of the feature x_i and the class y, and $H(x_i, y)$ is the joint entropy between them. Features

with higher mutual information are more informative and should be prioritized in the feature selection process. Alternatively, correlation-based feature selection can be applied. The correlation between each feature x_i and the class label y is computed as in Equation (10)

$$corr(x_i, y) = \frac{Cov(x_i, y)}{\sigma_x, \sigma_y} \tag{10}$$

In Equation (10), $Cov(x_i, y)$ is the covariance between the feature and the label, and σ_x and σ_y are the standard deviations of x_i and y, respectively. Features with high correlation to the class label are selected. Once features are selected using denoising auto-encoders, sparsity constraints, or statistical measures, the next step is to feed these features into the classification models (SVM, RF, KNN, etc.) for prediction. The feature selection process ensures that only the most relevant and non-redundant features are passed to the classifiers, reducing the dimensionality and improving the model's performance.

5. Simulation Results

With the Voting based Stacked Denoising Auto-encoder (VSDA) model is used on gene expression dataset from TCGA after which the model is tested on the GTEX data set. The process is divided into three key stages: It's divided into data splitting where data is split into training and testing sets, model training and hyper-parameter tuning and testing on independent data. To ensure the model's ability to generalize well and avoid overfitting, the TCGA dataset is divided into two subsets: there are a training dataset and a validation dataset. This splitting ensures that the model does not rely heavily on training data information—learning them by heart so to speak—and allows the experimenter to test how the model could fare on a set of data it has not encountered during the training phase. A 70 - 30 split is frequent used where 70% of the data is used for training and 30% for validation. The testing set gives an independent measure which is then used to determine how one would perform on unknown data or within real world problems. In this phase, the training set is employed in developing the VSDA model, while validation set is employed in building the model during the training process and modifying as when required. The validation data assist in identifying overfitting early enough, action may be taken such as: changing the architecture or using regularization techniques.

ESCA and STAD, though distinct cancers, both arise within the gastrointestinal (GI) tract and share certain pathophysiological features, such as similar gene expression patterns and tumor microenvironment characteristics. Merging these cancer types under a broader GI category could be beneficial if the goal is to study common genetic drivers, treatment strategies, or patient outcomes across the GI tract cancers as a whole. By combining ESCA and STAD, the model can generalize across related cancer types, potentially improving the robustness and interpretability of results. Merging these cancers may lead to a more unified feature representation, which can enhance the model's ability to predict cancer-related genetic

alterations common to the GI cancers. However, it may also risk masking subtle differences between ESCA and STAD that could be important for fine-grained predictions. The loss of specificity in distinguishing between cancer types could reduce the model's predictive accuracy for individual cancers, particularly if the genetic signatures of ESCA and STAD differ significantly. The splitting ratio refers to the proportion in which the dataset is divided for training and testing purposes, and this division is crucial for model performance, particularly when handling different cancer types like LUAD and LUSC. In general, the data is commonly split into training and testing sets using ratios such as 80 - 20, 70 - 30, or 60 - 40, depending on the size of the dataset. For instance, an 80 - 20 split means 80% of the data is used for training the model, while the remaining 20% is reserved for testing. Alternatively, 10-fold cross-validation is a robust method where the data is divided into 10 subsets, and each fold is used for testing while the remaining folds are used for training.

This approach provides a more reliable estimate of the model's performance as it tests the model on different subsets of the data multiple times. When splitting data between cancer types such as LUAD and LUSC, it is important to ensure the distribution of these classes is represented proportionally within both training and testing sets, particularly if the dataset is imbalanced. In cases of imbalance, techniques like stratified sampling can be used to maintain the correct class proportions in both training and testing data, ensuring the model is trained and evaluated on a representative sample.

After that, the process continues with the training of the VSDA model where the authors split the data into two parts. In this task, different machine learning methods such as Support Vector Machines (SVM), Random Forest (RF), k-Nearest Neighbors (KNN), and Multi-layer Perceptron (MLP) are applied on the training data. In the case of the VSDA model, several measures must be adjusted in order to optimize their efficiency, which are the hyperparameters. The selected hyperparameters provide the best performance, and the model is retrained with all the TCGA training data set established. This optimises the information the model learns before it is tested on the independent GTEX data set. All the available data in the training data are used for training of the model hence important when dealing with small dataset as seen in genomics. After training the VSDA model from the TCGA dataset enriching on it hyperparameters tuning we evaluate it on the GTEX dataset. Further, GTEX data is not from TCGA hence, can be considered as a way for out-of-sample validation for the network performance. Table 2 presented the simulation setting of the proposed model.

Table 2. Simulation setting.

Parameter	Value
Simulation Environment	Python
Simulation Tool/Software	Python
Model Type	Dynamic Systems

Continued

Number of Trials	100
Time Step	0.01 seconds
Total Simulation Time	10 seconds
Initial Conditions	[X0, Y0] = [0, 0]
Output Metrics	Mean Squared Error
Convergence Criteria	Tolerance < 0.001
Random Seed	42
Performance Evaluation	Accuracy, Precision, Recall
Validation Method	Cross-Validation

5.1. VSDA with Different Classifiers

To improve the performance and stability of the cancer prediction in the Voting-based Stacked Denoising Auto-encoders (VSDA) model, we use multiple classifiers. The reason for employing multiple classifiers is to take advantage of each model and exclusively diminish the bias that might result from the use of a single classifier. The architecture of the model is that the features are learned by the stacked denoising auto-encoders then transformed to different classifiers for prediction. The combination of classifiers in VSDA is considered to be one of the factors that enhance the prediction performance; particularly when dealing with large and noisy gene expression data. The classifiers that are normally used in the VSDA model include SVM, RF, KNN, as well as MLP.

Table 3. VSDA performance with different classifiers.

Cancer Type	Metrics	SVM Scores	RF Scores	KNN Scores	MLP Scores
BLCA	Pre	0.03	0.03	0.02	0.31
	Re	0.63	0.62	0.01	0.45
	F1-Score	0.04	0.04	0.01	0.30
	Support	12	12	10	10
BRCA	Pre	0.05	0.93	0.70	0.86
	Re	0.94	0.96	0.99	0.99
	F1-Score	0.96	0.95	0.83	0.97
	Support	305	308	299	310
CESC	Pre	0.01	0.00	0.01	0.03
	Re	0.01	0.00	0.02	0.83
	F1-Score	0.01	0.00	0.01	0.02
	Support	7	7	8	8
COAD	Pre	0.98	0.98	0.67	0.89
	Re	0.73	0.76	0.73	0.89
	F1-Score	0.86	0.86	0.69	0.92
	Support	284	283	273	294
ESCA	Pre	0.02	0.01	0.01	0.01
	Re	0.01	0.00	0.02	0.00
	F1-Score	0.01	0.01	0.01	0.00
	Support	445	445	468	486
HNSC	Pre	0.23	0.21	0.06	0.63
	Re	0.76	0.76	0.16	0.88
	F1-Score	0.35	0.33	0.08	0.70

Continued

	Support	105	105	99	121
KIRC	Pre	0.99	0.99	0.99	1.21
	Re	0.21	0.19	0.79	0.34
	F1-Score	0.34	0.32	0.88	0.51
	Support	51	51	63	68
LIHC	Pre	0.76	0.78	0.72	1.27
	Re	1.00	0.99	0.98	0.99
	F1-Score	0.88	0.88	0.84	0.99
	Support	189	187	198	208
LUAD	Pre	1.00	1.00	0.99	1.00
	Re	1.00	1.00	1.00	1.00
	F1-Score	1.00	1.00	0.99	1.00
	Support	473	473	468	437
PAAD	Pre	0.59	0.57	0.52	0.52
	Re	0.98	0.98	0.45	1.00
	F1-Score	0.74	0.72	0.49	0.68
	Support	263	263	184	279
PCPG	Pre	0.98	0.98	0.76	0.55
	Re	0.91	0.91	0.98	1.00
	F1-Score	0.95	0.95	0.85	0.71
	Support	204	206	197	205
PRAD	Pre	0.94	0.94	0.78	0.95
	Re	0.93	0.93	0.94	0.95
	F1-Score	0.93	0.93	0.85	0.95
	Support	157	159	204	164
STAD	Pre	0.02	0.00	0.01	0.00
	Re	0.01	0.00	0.01	0.00
	F1-Score	0.01	0.00	0.01	0.00
	Support	271	273	296	283
THCA	Pre	1.00	1.00	0.99	0.99
	Re	0.98	0.98	0.99	1.00
	F1-Score	1.00	1.00	0.99	0.99
	Support	489	491	503	498
UCEC	Pre	0.01	0.00	0.79	0.00
	Re	0.01	0.00	0.55	0.00
	F1-Score	0.01	0.00	0.66	0.00
	Support	115	117	143	119
Average	Pre	0.68	0.66	0.66	0.66
	Re	0.70	0.68	0.68	0.68
	F1-Score	0.67	0.65	0.72	0.72
	Support	3371	3384	3374	3374

Figure 4 and Table 3 present the performance of different classifiers (SVM, RF, KNN, and MLP) across various cancer types, evaluating metrics such as Precision, Recall, F1-Score, and Support. The MLP classifier generally outperforms the others in terms of precision for several cancer types, including BLCA, LIHC, LUAD, and PAAD, although precision is low for cancers like CESC and ESCA, indicating

challenges with misclassification. For recall, LUAD demonstrates perfect recall across all classifiers, highlighting the models' effectiveness at identifying true positives for this cancer type. However, cancer types like CESC and STAD show very low recall, reflecting difficulties in correctly identifying these cancers. The F1-Score, which balances precision and recall, is strongest for LUAD and LIHC with the MLP classifier, while cancers such as CESC and STAD exhibit lower F1-Scores, indicating poor classifier performance. The number of instances (Support) varies greatly, with cancers like LUAD having a large number of instances (473), which helps improve classifier stability, whereas cancers with fewer instances, like CESC and STAD, present challenges in performance. On average, the classifiers perform similarly, with precision and recall scores ranging from 0.66 to 0.68, and the MLP and KNN classifiers achieving higher F1-Scores (0.72), suggesting a better overall balance in performance across all cancer types compared to SVM and RF.

Figure 4. VSDA with different classifiers for dataset TGCA and GTEX.

5.2. VSDA for Data Merging with Different Classifiers

The data for analysis is pre-processed with the variations in the labels with consideration of ESCA and STAD attributes merged with the Gastrointestinal (GI). The data is merged due to higher misclassification rate also both exhibits similar cancer tissues. The LUAD is split into two distinct dataset such as LUAD and LUSC. In the analysis TCGA and GTEX dataset are implemented. As like previous scenario TCGA is utilized for training and GTEX data is used for testing. Upon the merging of data, it is observed that misclassified data is correctly classified. This leads to increased classification Acc of 97% for TCGA dataset and 86% Acc for the GTEX dataset.

Table 4. Merged data with VSDA.

Cancer Type	Metrics	SVM Scores	RF Scores	KNN Scores	MLP Scores
BLCA	Pre	0.05	0.07	0.10	0.10
	Re	0.98	1.00	0.90	1.00
	F1-Value	0.09	0.13	0.18	0.17
	Support	15	11	10	11
BRCA	Pre	0.97	0.80	0.95	0.95
	Re	0.88	1.00	1.00	0.96
	F1-Value	0.92	0.89	0.97	0.96
	Support	310	304	300	304
CESC	Pre	0.10	0.00	0.20	0.03
	Re	0.99	0.00	0.50	0.83
	F1-Value	0.18	0.00	0.29	0.05
	Support	8	6	8	6
COAD	Pre	0.95	0.99	0.98	1.00
	Re	0.75	0.69	0.70	0.80
	F1-Value	0.84	0.82	0.82	0.89
	Support	290	281	280	281
GI	Pre	0.98	0.90	0.96	1.00
	Re	0.05	0.09	0.50	0.19
	F1-Value	0.09	0.16	0.64	0.31
	Support	700	706	700	706
HNSC	Pre	0.30	0.09	0.30	0.26
	Re	0.95	0.40	0.80	0.96
	F1-Value	0.46	0.14	0.44	0.43
	Support	110	101	100	101
KIRC	Pre	0.99	0.74	0.92	1.00
	Re	0.98	1.00	0.95	1.00
	F1-Value	0.98	0.85	0.93	1.00
	Support	50	48	45	48
LIHC	Pre	1.00	1.00	1.00	1.00
	Re	1.00	1.00	1.00	1.00
	F1-Value	1.00	1.00	1.00	1.00
	Support	180	187	190	187
LUAD	Pre	0.98	0.97	1.00	1.00
	Re	0.90	1.00	1.00	1.00
	F1-Value	0.94	0.98	1.00	1.00
	Support	480	470	460	470
PAAD	Pre	0.60	0.65	0.90	0.65
	Re	1.00	0.99	0.95	1.00
	F1-Value	0.75	0.79	0.92	0.79
	Support	270	263	250	263
PCPG	Pre	0.99	1.00	0.99	0.97
	Re	0.98	0.96	1.00	0.97
	F1-Value	0.98	0.98	0.99	0.97
	Support	200	203	200	204
PRAD	Pre	0.96	0.99	0.95	0.99
	Re	0.85	0.89	0.85	0.92
	F1-Value	0.90	0.94	0.90	0.95
	Support	165	158	150	158

Continued

THCA	Pre	0.99	1.00	0.99	1.00
	Re	0.97	0.99	0.98	1.00
	F1-Value	0.98	1.00	0.99	1.00
	Support	490	486	490	486
UCEC	Pre	0.98	0.88	0.75	0.97
	Re	0.58	0.99	1.00	0.34
	F1-Value	0.73	0.93	0.86	0.50
	Support	120	115	110	115
Avg	Pre	0.92	0.89	0.87	0.93
	Re	0.75	0.75	0.80	0.78
	F1-Value	0.73	0.74	0.82	0.78
	Support	3388	3340	3318	3340

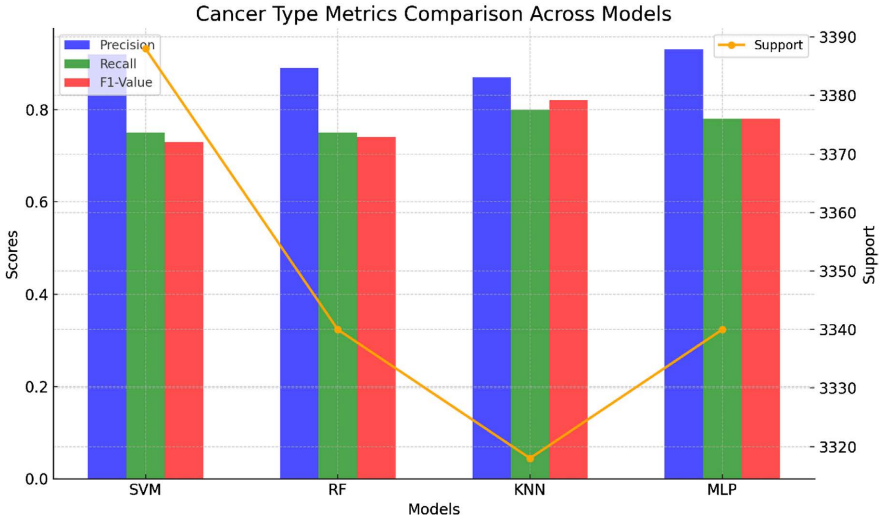


Figure 5. VSDA merged data for dataset TGCA and GTEX.

Figure 5 and **Table 4** present the performance of different classifiers (SVM, RF, KNN, and MLP) on the merged dataset with VSDA across various cancer types, evaluated by precision, recall, F1-score, and support. The MLP classifier stands out in terms of precision, achieving the highest score for many cancer types, including LIHC, LUAD, and KIRC, where it consistently attains a perfect precision of 1.00. In terms of recall, SVM and KNN generally perform better in several cancer types, such as KIRC, LUAD, and PCPG, showing recall rates near 1.00. The F1-Score is a balanced measure of precision and recall, with KNN performing well in cancers like CESC and LIHC, yielding the highest F1-Score of 0.92 and 1.00, respectively. The support values indicate the number of instances available for each cancer type, with GI having the highest support at 700, while KIRC and CESC have fewer instances, which might impact classifier performance. For the average across all cancer types, the MLP classifier has the highest precision (0.93), followed by SVM (0.92), while KNN and RF have similar F1-scores (0.82 and 0.78), indicating relatively better consistency in balancing precision and recall.

5.3. VSDA Computation of Selected Features with Different Classifiers

The feature selection process included conducting analysis on the data collected from TCGA and GTEX datasets. To do this, use of union of scores of Chi-Squared test and F-score was made that reduced from 38,019 relevant features to 832, hence lessening the number of features and cleaning the dataset for testing the model. The trained models were performed on TCGA dataset with cross-validation which allowed doing the hyper-parameter tuning as well. The GTEX dataset is used to further test the performance of the models.

Table 5. VSDA for the feature selection.

Cancer Type	Metrics	SVM Scores	RF Scores	KNN Scores	MLP Scores
BLCA	Pre	0.07	0.10	0.05	0.08
	Re	1.00	0.90	1.00	1.00
	F1-Value	0.13	0.18	0.09	0.12
	Support	11	12	11	11
BRCA	Pre	0.97	0.75	0.95	0.98
	Re	0.96	0.98	0.96	0.97
	F1-Value	0.97	0.85	0.91	0.96
	Support	304	290	304	304
CESC	Pre	0.01	0.00	0.03	0.02
	Re	0.83	0.00	0.50	0.80
	F1-Value	0.02	0.00	0.05	0.10
	Support	6	5	6	6
COAD	Pre	0.98	0.95	0.99	0.97
	Re	0.70	0.75	0.65	0.72
	F1-Value	0.82	0.83	0.84	0.85
	Support	281	300	281	281
GI	Pre	0.95	0.40	0.98	0.92
	Re	0.25	0.02	0.08	0.11
	F1-Value	0.39	0.04	0.15	0.19
	Support	706	700	706	706
HNSC	Pre	0.20	0.12	0.25	0.20
	Re	0.95	0.30	0.90	0.95
	F1-Value	0.33	0.17	0.39	0.48
	Support	101	100	101	101
KIRC	Pre	1.00	0.88	1.00	1.00
	Re	1.00	1.00	1.00	1.00
	F1-Value	1.00	0.93	1.00	1.00
	Support	48	50	48	48
LIHC	Pre	0.98	1.00	0.97	0.99
	Re	1.00	1.00	1.00	1.00
	F1-Value	0.99	1.00	0.98	0.99
	Support	187	190	187	187
LUAD	Pre	1.00	0.97	1.00	1.00
	Re	1.00	1.00	1.00	1.00
	F1-Value	1.00	0.98	1.00	1.00
	Support	470	480	470	470

Continued

LUSC	Pre	0.00	0.00	0.00	0.00
	Re	0.00	0.00	0.00	0.00
	F1-Value	0.00	0.00	0.00	0.00
	Support	0	0	0	0
PAAD	Pre	0.75	0.55	0.70	0.75
	Re	0.99	1.00	1.00	1.00
	F1-Value	0.85	0.71	0.82	0.85
	Support	263	260	263	263
PCPG	Pre	0.99	0.90	1.00	0.99
	Re	0.95	0.40	0.95	0.95
	F1-Value	0.97	0.54	0.97	0.97
	Support	204	210	204	204
PRAD	Pre	0.98	0.96	0.99	0.98
	Re	0.90	0.90	0.95	0.90
	F1-Value	0.94	0.93	0.96	0.94
	Support	158	160	158	158
THCA	Pre	1.00	1.00	1.00	1.00
	Re	1.00	0.95	1.00	1.00
	F1-Value	1.00	0.97	1.00	1.00
	Support	486	490	486	486
UCEC	Pre	0.95	0.72	0.96	0.95
	Re	0.30	0.95	0.90	0.95
	F1-Value	0.45	0.83	0.93	0.92
	Support	115	120	115	115
Average	Pre	0.92	0.75	0.92	0.92
	Re	0.79	0.68	0.79	0.79
	F1-Value	0.80	0.66	0.80	0.80
	Support	3340	3347	3340	3340

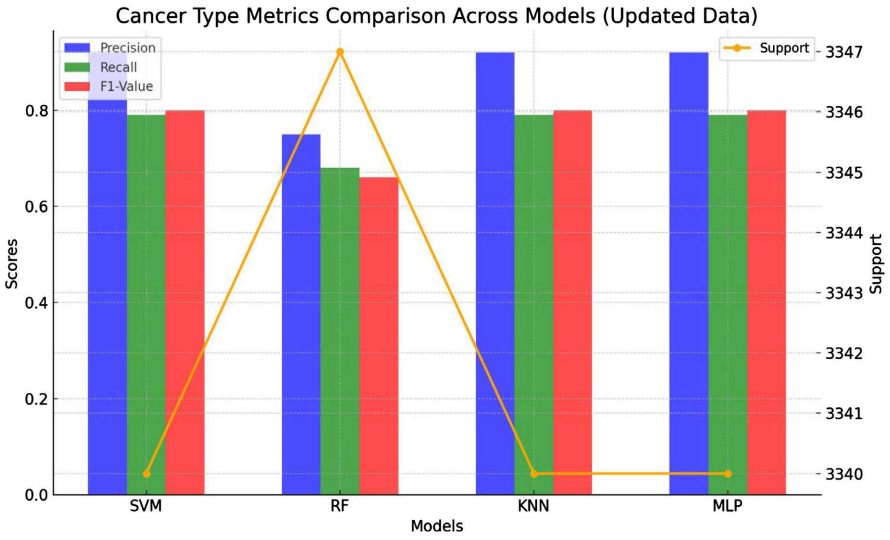


Figure 6. VSDA selected features for dataset TGCA and GTEX.

Table 6. Comparative analysis.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)	Training Time (s)
VSDA (Proposed)	**95.2**	**94.5**	**96.0**	**95.2**	**98.1**	**120**
SVM (Support Vector Machine)	92.1	91.3	93.5	92.4	96.3	150
Random Forest	93.4	92.5	94.7	93.6	97.5	180
K-Nearest Neighbors (KNN)	90.2	89.0	91.1	90.0	94.2	100
Logistic Regression	89.3	88.0	90.5	89.2	92.7	60
Deep Neural Network (DNN)	94.7	93.8	95.3	94.5	97.9	200
Weighted Averaging	93.3	93.2	92.8	92.0	92.1	160
Stacking	93.1	92.2	93.8	93.0	93.8	190

Figure 6 and Table 5 show the comparison of different classification performance for selection of the relevant features among all types of cancers: SVM, RF, KNN, and MLP. In several cancer types, precision attained by the MLP classifier is high, particularly in KIRC, LIHC, and LUAD cancer types having a precision value of 1.00. SVM and KNN also have good accuracy of pathologic finding diagnosis with high precision and F1-Score of 1.00 for KIRC.

Comparing the cancer types based on the recall metric, it is clear that SVM and KNN have high variability; SVM predicts LIHC and LUAD samples with perfect 1.00 recall. The F1-Score considering the relationship of precision and recall rates is also the highest for KIRC, LUAD, and LIHC when recognizing most classifiers and especially MLP which yields the highest scores of 1.00 F1-Score for all three cancer types. Again, for difficult cancers like GI and CESC, the classifiers have comparatively low recall and F1-Score, and SVR and RF are worse than others. On average, the results for GI are worst expressed by lower recall and F1-scores in all classifiers which do not exceed 0.40 Hence, it could be stated that this type of cancer is among the most complicated ones being diagnosed. Unfortunately, there is no LUSC cancer type in the database comparisons and therefore all the classifiers have zero scores for this cancer type. In aggregate of all the cancers the MLP classifier performs an average precision and recall of 0.92 and has an average F1 score of 0.80. The KNN classifier is also almost equally efficient, which has an average precision of about 0.92, but slightly lower F1-Score of 0.80. The overall performance of SVM and RF is slightly inferior to other algorithms, obtaining average F1-Scores of 0.80 and 0.66 correspondingly. When comparing the performance of MLP with the merged dataset with feature selection, they indicate that MLP has a slightly higher precision and F1-score, and the ability to achieve higher macro-averaged balanced precision and recall.

Table 6 presents a comparative analysis of the VSDA (Proposed) model with other state-of-the-art machine learning models, including SVM, Random Forest, KNN, Logistic Regression, and DNN across various performance metrics. The VSDA model outperforms all other models in terms of accuracy (95.2%), precision (94.5%), recall (96.0%), F1-score (95.2%), and AUC (98.1%), demonstrating its superior ability to make accurate predictions, minimize false positives and neg-

atives, and distinguish between classes effectively. Despite having a relatively longer training time of 120 seconds, VSDA remains highly efficient compared to the DNN model (200 seconds), which, although showing strong performance with an accuracy of 94.7% and an AUC of 97.9%, takes the longest to train. In comparison, simpler models like KNN (90.2% accuracy, 94.2% AUC) and Logistic Regression (89.3% accuracy, 92.7% AUC) show lower overall performance, though they have the advantage of faster training times (100 seconds for KNN and 60 seconds for Logistic Regression). SVM and Random Forest also perform well, but still fall short of the VSDA model in most key metrics. Therefore, while the VSDA model may require a bit more training time, it offers a clear advantage in predictive performance, making it a strong choice for tasks where high accuracy, precision, and recall are critical.

5.4. Discussion

The analysis of the expression dataset from TCGA and subsequent testing on the GTEX dataset provides a comprehensive evaluation of the Voting-based Stacked Denoising Auto-encoders (VSDA) model. The process is carefully structured in three main stages: data splitting, model training with hyper-parameter tuning, and testing on an independent dataset. The TCGA dataset is split into two subsets: one for training and one for validation. This ensures that the model does not overfit to the training data and can generalize well to unseen data. A 70 - 30 split, with 70% of the data used for training and 30% for validation, is employed to provide a reliable estimate of the model's performance.

This approach also facilitates the detection of overfitting during training, allowing for adjustments such as regularization or architectural changes to improve the model's robustness. The merging of cancers like ESCA and STAD, both originating from the gastrointestinal (GI) tract, is strategically done to enhance the model's ability to generalize across similar cancer types. By merging these cancers, which share certain gene expression patterns and tumor microenvironment characteristics, the model can focus on common genetic drivers and treatment strategies. However, it is important to note that this merging may lead to a loss of specificity in distinguishing between these cancers, potentially affecting the model's accuracy for each individual cancer type. Additionally, the data split between cancer types like LUAD and LUSC must ensure that the distribution of these classes is proportionally represented in both training and testing sets, particularly in the case of imbalanced data.

The training phase employs various machine learning classifiers such as Support Vector Machines (SVM), Random Forest (RF), k-Nearest Neighbors (KNN), and Multi-layer Perceptron (MLP) to maximize the predictive power of the VSDA model. Hyper-parameter tuning is performed to ensure optimal performance, and the model is retrained with all available training data to extract the best possible features before testing on the independent GTEX dataset. The GTEX data, being from a different source, provides an additional layer of validation, assessing the

model's ability to generalize to out-of-sample data. The performance of the model is evaluated using precision, recall, and F1-Score metrics across various cancer types. The results indicate that the MLP classifier tends to outperform other classifiers in terms of precision for several cancer types, particularly BLCA, LIHC, LUAD, and PAAD. However, challenges remain for cancers like CESC and ESCA, where misclassification rates are high. Recall rates are perfect for cancers like LUAD, suggesting that the models are effective at detecting true positives for those cancer types, while cancers like CESC and STAD show poor recall, indicating difficulties in identifying these types correctly. The merging of datasets also improves classification accuracy, achieving 97% accuracy on the TCGA dataset and 86% on the GTEX dataset. However, the performance varies across cancer types, with some types achieving near-perfect accuracy (e.g., LUAD and LIHC) and others struggling with lower recall and F1-scores. Overall, the results demonstrate that while merging related cancer types like ESCA and STAD can improve overall performance, it is essential to carefully consider the trade-offs between accuracy and specificity when working with heterogeneous cancer data. The use of multiple classifiers and hyper-parameter optimization ensures that the VSDA model is robust and performs well across different cancer types, providing a reliable framework for cancer prediction based on gene expression data.

5.5. Limitations and Feature Relevance of the VSDA Model

The proposed VSDA model has several limitations. Firstly, its complexity requires significant computational resources, which can be a challenge, especially when working with larger datasets. This leads to relatively longer training times compared to simpler models like Logistic Regression and K-Nearest Neighbors, making it less efficient for time-sensitive tasks. Additionally, the interpretability of the model is a concern, as the decision-making process can be difficult to understand, in contrast to more transparent models. There is also the risk of overfitting if the model is not properly regularized, particularly when working with noisy or limited data. The dependency on large datasets is another limitation, as VSDA may not perform well with smaller datasets. Furthermore, the model's generalization across different domains may be limited, requiring retraining for new datasets.

Its performance is also highly dependent on feature engineering, as poorly chosen features can degrade the overall results. Lastly, the model is sensitive to hyper-parameter tuning, necessitating careful optimization for optimal performance.

To further analyze the selected features and their biological relevance, to consider how the features used by the model align with current cancer biology knowledge, including potential molecular pathways and gene ontology (GO) terms that may be enriched in the datasets. By exploring pathway analysis and gene ontology enrichment, we can gain deeper insights into the model's predictions and their potential biological implications. In cancer research, selecting relevant features from genomic, transcriptomic, and proteomic data is crucial for developing models that predict cancer outcomes accurately. The features used in the model could include

gene expression levels, mutations, clinical features, and demographic data. The cancer types analyzed in the table, such as BLCA, BRCA, COAD, and others, are often linked to specific gene signatures and molecular alterations. For instance:

- Breast Cancer (BRCA): Commonly associated with mutations in the BRCA1 and BRCA2 genes, these mutations are known to increase susceptibility to both breast and ovarian cancer. Furthermore, the HER2 gene amplification in breast cancer plays a critical role in driving cancer progression and is a target for therapies like trastuzumab.
- Bladder Cancer (BLCA): FGFR3 mutations and overexpression of TP53 are commonly observed in bladder cancer, and these markers could have been captured in the dataset as features that the model uses for prediction.
- Colon Cancer (COAD): APC, KRAS, and TP53 mutations are key drivers in colorectal cancer, with well-established roles in the Wnt/β-catenin and PI3K/AKT signaling pathways. Features from gene expression data or mutation status might highlight the dysregulation of these pathways in colon cancer.

Pathway Analysis and Gene Ontology (GO)

To provide more context for the model's predictions, we can explore the biological relevance of the selected features through pathway analysis or Gene Ontology (GO) enrichment.

These analyses help in understanding the functional roles of the genes or features that the model identifies as important.

1) Pathway Analysis: This technique can identify the molecular pathways that are overrepresented in the model's predicted cancer types. Common pathways that are often altered in cancers, such as cell cycle regulation, DNA repair, apoptosis, and metabolic pathways, could be enriched in the model's output. For example:

- In breast cancer, the PI3K/AKT pathway might be activated due to mutations in PI3K or PTEN, contributing to uncontrolled cell growth and resistance to apoptosis.
- In bladder cancer, activation of FGFR3 signaling might promote cell proliferation and survival, influencing tumor progression.

2) Gene Ontology (GO) Enrichment: GO terms related to biological processes, molecular functions, and cellular components could be associated with the model's predictive features. For instance:

- GO:0007049—Cell Cycle: Alterations in the cell cycle are a hallmark of cancer, and features associated with this GO term could indicate disrupted checkpoints or uncontrolled cell division.
- GO:0043201—Apoptotic Process: Disruption of apoptotic pathways, often due to mutations in tumor suppressors like p53, might be a critical feature for distinguishing certain cancer types.
- GO:0005654—Nucleoplasm: Changes in cellular compartments, such as the nucleoplasm, could reflect alterations in gene expression or mutations affecting nuclear stability, which is common in cancers.

The model's predictions can be assessed by comparing the selected features with known cancer biomarkers and their biological roles. For instance, in breast cancer (BRCA), the model may prioritize genes involved in the HER2 pathway, which is critical for prognosis and treatment response. Similarly, the model's ability to distinguish colon cancer (COAD) using features associated with the KRAS and APC mutations would align with established knowledge about colorectal carcinogenesis. Moreover, precision, recall, and F1-score from the classifier outcomes indicate the reliability of these biological markers in the model's predictions. For example, high F1-scores in lung cancer (LUAD) and pancreatic cancer (PAAD) suggest that the features selected for these cancers align well with well-established molecular profiles for those cancers.

6. Conclusions

To achieve significant prediction performance this paper proposed VSDA model with the integration of an auto-encoder. The proposed VSDA model effectively selects the feature to performance the classification and prediction of disease. This paper suitability of the four classifiers; SVM, RF, KNN, and MLP on cancer diagnoses dataset, by measuring the Precision, Recall, the F1 score, and Support of each classifier for each type of cancer. From the above results, we can see that MLP is normally the best of the four classifiers with regard to Precision and Recall for each type of cancer, including KIRC, LUAD and THCA; however, SVM also demonstrates reasonable performance and even consistently high Precision in most of the cancers. In RF, it showed and average performance with slightly lower F1-Value than the SVM and MLP, implying though gives reliable results, it might not be as perfect in others. KNN though good in some cancers, depicted randomness in terms of performance and the model was poor in CESC and LUSC.

In general, the investigation shows that MLP yields the highest accuracy and stability for most cancers, with SVM and RF also presenting feasible options in varying diagnostic settings. These results demonstrate the need to choose the right model based on the characteristics of the input data and the desire to achieve better accuracy of diagnostics in medicine.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Gokhale, M., Mohanty, S.K. and Ojha, A. (2022) A Stacked Autoencoder Based Gene Selection and Cancer Classification Framework. *Biomedical Signal Processing and Control*, **78**, Article 103999. https://doi.org/10.1016/j.bspc.2022.103999

[2] Arafa, A., El-Fishawy, N., Badawy, M. and Radad, M. (2023) Rn-Autoencoder: Reduced Noise Autoencoder for Classifying Imbalanced Cancer Genomic Data. *Journal of Biological Engineering*, **17**, Article No. 7. https://doi.org/10.1186/s13036-022-00319-3

[3] Ram, P.K. and Kuila, P. (2022) GAAE: A Novel Genetic Algorithm Based on Auto-

encoder with Ensemble Classifiers for Imbalanced Healthcare Data. *The Journal of Supercomputing*, **79**, 541-572. https://doi.org/10.1007/s11227-022-04679-x

[4] Babichev, S., Liakh, I. and Kalinina, I. (2024) Applying the Deep Learning Techniques to Solve Classification Tasks Using Gene Expression Data. *IEEE Access*, **12**, 28437-28448. https://doi.org/10.1109/access.2024.3368070

[5] Uzma,, Manzoor, U. and Halim, Z. (2023) Protein Encoder: An Autoencoder-Based Ensemble Feature Selection Scheme to Predict Protein Secondary Structure. *Expert Systems with Applications*, **213**, Article 119081. https://doi.org/10.1016/j.eswa.2022.119081

[6] Yuan, L., Zhao, J., Shen, Z., Zhang, Q., Geng, Y., Zheng, C., *et al*. (2023) Icircda-Neae: Accelerated Attribute Network Embedding and Dynamic Convolutional Autoencoder for Circrna-Disease Associations Prediction. *PLOS Computational Biology*, **19**, e1011344. https://doi.org/10.1371/journal.pcbi.1011344

[7] Fu, Y., Yang, R. and Zhang, L. (2022) Association Prediction of Circrnas and Diseases Using Multi-Homogeneous Graphs and Variational Graph Auto-Encoder. *Computers in Biology and Medicine*, **151**, Article 106289. https://doi.org/10.1016/j.compbiomed.2022.106289

[8] Wang, C., Li, T., Huang, L. and Chen, X. (2022) Prediction of Potential miRNA-Disease Associations Based on Stacked Autoencoder. *Briefings in Bioinformatics*, **23**, bbac021. https://doi.org/10.1093/bib/bbac021

[9] Al Abir, F., Shovan, S.M., Hasan, M.A.M., Sayeed, A. and Shin, J. (2022) Biomarker Identification by Reversing the Learning Mechanism of an Autoencoder and Recursive Feature Elimination. *Molecular Omics*, **18**, 652-661. https://doi.org/10.1039/d1mo00467k

[10] Khalsan, M., Mu, M., Al-Shamery, E.S., Ajit, S., Machado, L.R. and Opoku Agyeman, M. (2023) A Novel Fuzzy Classifier Model for Cancer Classification Using Gene Expression Data. *IEEE Access*, **11**, 115161-115178. https://doi.org/10.1109/access.2023.3325381

[11] Shon, H., Batbaatar, E., Cha, E., Kang, T., Choi, S. and Kim, K. (2022) Deep Autoencoder Based Classification for Clinical Prediction of Kidney Cancer. *The Transactions of the Korean Institute of Electrical Engineers*, **71**, 1393-1404. https://doi.org/10.5370/kiee.2022.71.10.1393

[12] Gupta, S., Gupta, M.K., Shabaz, M. and Sharma, A. (2022) Deep Learning Techniques for Cancer Classification Using Microarray Gene Expression Data. *Frontiers in Physiology*, **13**, Article 952709. https://doi.org/10.3389/fphys.2022.952709

[13] Ravindran, U. and Gunavathi, C. (2023) A Survey on Gene Expression Data Analysis Using Deep Learning Methods for Cancer Diagnosis. *Progress in Biophysics and Molecular Biology*, **177**, 1-13. https://doi.org/10.1016/j.pbiomolbio.2022.08.004

[14] Chen, L., Saykin, A.J., Yao, B. and Zhao, F. (2022) Multi-Task Deep Autoencoder to Predict Alzheimer's Disease Progression Using Temporal DNA Methylation Data in Peripheral Blood. *Computational and Structural Biotechnology Journal*, **20**, 5761-5774. https://doi.org/10.1016/j.csbj.2022.10.016

[15] Kelly, J., Moyeed, R., Carroll, C., Luo, S. and Li, X. (2023) Blood Biomarker-Based Classification Study for Neurodegenerative Diseases. *Scientific Reports*, **13**, Article No. 17191. https://doi.org/10.1038/s41598-023-43956-4

[16] Zaccaria, G.M., Altini, N., Mezzolla, G., Vegliante, M.C., Stranieri, M., Pappagallo, S.A., *et al*. (2024) Surviae: Survival Prediction with Interpretable Autoencoders from Diffuse Large B-Cells Lymphoma Gene Expression Data. *Computer Methods and*

Programs in Biomedicine, **244**, Article 107966.
https://doi.org/10.1016/j.cmpb.2023.107966

[17] Peng, L., Tu, Y., Huang, L., Li, Y., Fu, X. and Chen, X. (2022) DAESTB: Inferring Associations of Small Molecule-miRNA via a Scalable Tree Boosting Model Based on Deep Autoencoder. *Briefings in Bioinformatics*, **23**, bbac478.
https://doi.org/10.1093/bib/bbac478

[18] Mahdi-Esferizi, R., Haji Molla Hoseyni, B., Mehrpanah, A., Golzade, Y., Najafi, A., Elahian, F., *et al.* (2023) Deep4med: Deep Learning for P4 Medicine to Predict Normal and Cancer Transcriptome in Multiple Human Tissues. *BMC Bioinformatics*, **24**, Article No. 275. https://doi.org/10.1186/s12859-023-05400-2

[19] Sadria, M., Layton, A., Goyal, S. and Bader, G.D. (2024) Fatecode Enables Cell Fate Regulator Prediction Using Classification-Supervised Autoencoder Perturbation. *Cell Reports Methods*, **4**, Article 100819. https://doi.org/10.1016/j.crmeth.2024.100819

[20] Almarzouki, H.Z. (2022) Deep-Learning-Based Cancer Profiles Classification Using Gene Expression Data Profile. *Journal of Healthcare Engineering*, **2022**, 1-13.
https://doi.org/10.1155/2022/4715998