


Investigation of Machine Learning on Gene Expression Data for Cancer Detection

Omar Abdul Razzaq 

The Information Department, The University of Qom, Qom 3716146611, Iran

Corresponding Author Email: aap6798@gmail.com



Copyright: ©2024 The author. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/mmep.110910>

ABSTRACT

Received: 3 February 2024

Revised: 31 May 2024

Accepted: 10 June 2024

Available online: 29 September 2024

Keywords:

cancer prediction, gene expression data, artificial intelligence, machine learning, deep learning, semi-supervised technique

Cancer remains a leading cause of global mortality due to delayed diagnoses and inadequate treatments from uncontrolled cell growth. Leveraging machine learning techniques can aid in early cancer prediction given available data. This study aims to improve tumor classification accuracy and efficiency based on gene expression patterns using deep learning algorithms. The primary approach involves constructing a feed-forward network (FFN) for binary classification, distinguishing between cancerous and healthy samples using the Cancer Genome Atlas (TCGA) database. Breast cancer, with ample samples in TCGA, and kidney cancer, with high mortality rates, were chosen for this study. Three feature extraction methods—Principal Component Analysis (PCA), Analysis of Variance (ANOVA), and Random Forests—were employed for preprocessing. The FFN achieved the highest accuracy for the kidney dataset using PCA with 300 principal components, yielding optimal accuracy and low error rates. For the breast dataset, PCA also produced favorable results, though requiring more principal components to retain sufficient variance. Comparative analysis showed PCA excelled in preserving variance and optimizing accuracy, with ANOVA also performing well, especially in the breast dataset, whereas Random Forests were less effective overall. These results highlight the importance of tailoring feature extraction methods and model architectures to specific dataset characteristics for the most accurate and efficient predictive models. This study demonstrates the potential of optimizing these parameters to enhance tumor classification model accuracy and reliability, providing valuable insights for improving diagnostic and treatment approaches in breast and kidney cancers.

1. INTRODUCTION

Bioinformatics is the field of using computational tools and algorithms for storing, manipulating, and retrieving significant information from biological data. Since 1980s, bioinformatics has undergone many changes in definition, such as a mundane one of “the use of computers to retrieve, process, analyze, and simulate biological information”, to specific one such as “the application of information science to biology, medicine, and life sciences” [1]. This field has rapidly grown into a wide research area with many different categories of research such as genomics, proteomics, systems biology, modeling and simulation, data mining, big data, networks and analysis, evolutionary computing, statistics and probability, and others [2, 3]. The word cancer is derived from the Greek word “karkinos” which means crab [1]. Greek physician Hippocrates first time used cancer word to describe the tumors. According to Hippocrates observation the basic part of the tumor appears like a crab body and the several extensions of the tumor seem to be the legs and claws of the crab [4]. Cancer represents a perilous ailment stemming from irregular cell division and the unrestrained proliferation of cells. Typically, cancerous cells exhibit distinctive behavior compared to normal cells and possess the capability to disseminate to

distant body regions. The dissemination of cancer cells to other bodily areas is known as metastasis [5]. The genesis of cancer initiates from the transformation of regular cells into malignant cancer cells, a multifaceted process typically advancing from pre-cancerous cells to the formation of malignant tumors. These alterations occur due to the interplay between an individual's genetic attributes and three primary external factors. Early identification and classification of cancer subtypes hold immense significance in providing improved diagnostic measures for patients. Consequently, predicting cancer subtypes (classes) during the initial stages has emerged as a crucial focal point in the realm of machine learning and medical science, garnering widespread attention from researchers and scientists worldwide.

RNA sequencing (RNA-seq) can detect cellular changes and analyze gene expression patterns within RNA, offering insights into the transcriptome. With available data, machine learning techniques can aid in early cancer prediction. RNA-seq helps researchers understand tumor classification and progression by monitoring gene expression and transcriptome changes in cancer RNA-seq data [6].

Several approaches have been proposed and investigated for classifying RNA-seq data. Recently, researchers have focused on quantile-transformed quadratic discriminant analysis for

high-dimensional RNA-seq data. They introduced a new classification method based on a model where the counts are marginally negative binomial but dependent. To select genes for classification, they first filtered out genes with low expression and conducted a likelihood ratio test (LRT).

There exist different clinical approaches to the diagnosis of cancer, which are described below. This paper presents two distinct approaches devised to tackle the binary classification challenge using TCGA mRNA sequencing data from breast and kidney cancer. Both approaches leverage FFN network structure for training. Furthermore, the study incorporates three diverse feature extraction algorithms PCA, ANOVA, and random forest as part of the pre-processing phase to assess their respective impacts on the outcomes. These methods collectively aim to discern the most effective strategy for handling and processing the data, shedding light on the significance of different feature extraction techniques in enhancing the classification performance of the networks [6].

2. RELATED WORK

Cancer, a feared illness initiated by genetic mutations, triggers uncontrolled and aberrant cell growth, manifesting as tumors in the initial stages that can swiftly metastasize to other body regions. This infection represents a basic worldwide wellbeing challenge. As per the Worldwide Disease Occurrence, Mortality, and Predominance (GLOBOCAN) project [7], an expected 8.2 million passings were credited to malignant growth in 2012 around the world. Consequently, early-stage cancer prediction has emerged as a pivotal focus area among scientists and researchers globally. Traditional cancer prediction methods rely on costly clinical diagnoses and tumor morphology analysis, which can be inaccurate and time-consuming [8]. Conventional approaches are frequently constrained by their reliance on expert assessments for tumor identification and their struggle to distinguish between various cancer subtypes. To overcome these challenges and offer a cost-effective initial diagnosis and prediction of cancer, contemporary computational techniques, such as microarray data analysis [9], have been employed. Microarray technology records thousands of concurrent gene expression profiles. The vast number of genes in microarray data far exceeds the available samples [10], leading to the presence of multiple ambiguous, overlapping, and indistinct cancer subtypes within the data [11]. Thus, developing classifiers equipped for accomplishing high prescient precision in characterizing malignant examples becomes basic [12]. Customary administered AI calculations [13], normally utilized for disease grouping using microarray quality articulation information, depend intensely on named tests to anticipate unlabeled ones. While gathering unlabeled quality articulation designs is somewhat clear, acquiring named tests is frequently costly, tedious, or testing. Accordingly, the shortage of marked examples habitually confines the relevance of conventional administered techniques for malignant growth arrangement and expectation.

In situations where natural information is obliged because of the shortage of clinically marked examples, utilizing dynamic learning as well as semi-managed strategies becomes instrumental in accomplishing uplifted exactness in malignant growth forecast. On the other hand, semi-supervised learning methods [14] capitalize on the distribution of unlabeled samples. This technique involves computationally selecting

'high confidence' unlabeled patterns, along with their predicted labels, from the unlabeled dataset. In recent studies on early cancer detection, various conventional machine learning methodologies, such as Support Vector Machines (SVM), k-Nearest Neighbor (kNN), Naïve Bayes (NB), Random Forest (RF), and their counterparts, have been extensively employed by Zhang et al. [15]. For instance, in a study by Sathe et al. [16], a genome based SVM strategy was proposed genome-based sarcoma classification. By employing Student's t-test, the authors selected 256 genes and used them to train a linear SVM classifier. The classifier successfully distinguished melanoma and soft tissue sarcoma, achieving high accuracy in leave-one-out cross-validation with 75 out of 76 instances correctly identified. To further enhance the performance of these machine learning methods, feature selection techniques have been incorporated. For example, in another study by Gunavathi et al. [17], SVM was combined with recursive feature elimination (SVM-RFE-PO). This approach utilized grid search and Partial Swarm Optimization for feature selection and a genetic algorithm for parameter tuning. The resulting model was able to identify a robust set of significant features for cancer classification. In a similar vein, Tabares-Soto et al. [18] deployed a Random Forest ensemble to extract 273 relevant genes while maintaining a robust classifier's predictability. Additionally, Li et al. [19] introduced a two-step feature selection strategy based on an attribute estimation method and Genetic Algorithm. Moreover, the Developmental Programming-prepared Help Vector Machine (EP-SVM) technique created by Mazlan et al. [20] used a probabilistic SVM way to deal with assess the results of double classifiers utilizing unmistakable class highlights. Overall, these studies demonstrate the use of various machine learning techniques, feature selection methods, and ensemble approaches to improve prediction accuracy in early cancer detection.

The related work presented in the manuscript underscores the significance of employing computational techniques, particularly machine learning algorithms, for early cancer detection and prediction using genomic data [21, 22]. Various conventional machine learning methodologies such as Support Vector Machines (SVM), k-Nearest Neighbor (kNN), Naïve Bayes (NB), and Random Forest (RF) have been extensively explored for cancer classification. These studies have demonstrated the effectiveness of integrating feature selection techniques with machine learning models to enhance prediction accuracy [23, 24].

The proposed research extends existing studies by focusing on the application of deep learning, specifically feed-forward networks (FFN) with supervised learning and ladder networks with semi-supervised learning, for tumor classification based on gene expression data. By leveraging deep learning techniques, the research aims to achieve more robust and accurate classification of cancerous and healthy samples. Additionally, the selection of breast cancer and kidney cancer datasets from the Cancer Genome Atlas (TCGA) database provides a comprehensive evaluation of the proposed methods on different cancer types, considering factors such as sample availability and mortality rates.

3. PROPOSED SYSTEM

This paper aims to perform predictive analyses using various models, aiming to classify provided datasets into specific cancer types. The main objective is to discern the

origin of cancerous tissue by analyzing their gene expression counts. This section encapsulates comprehensive information regarding the datasets employed in the current investigation, accompanied by details concerning the methods compared against the proposed approach. Figure 1 illustrates the proposed methodology diagram for tumor classification based on gene expression patterns using deep learning algorithms.

Subsequently, it highlights the assessment of performance evaluation measures. Lastly, a summary of the experimental setup is provided, encompassing the key aspects of the experimental configuration.

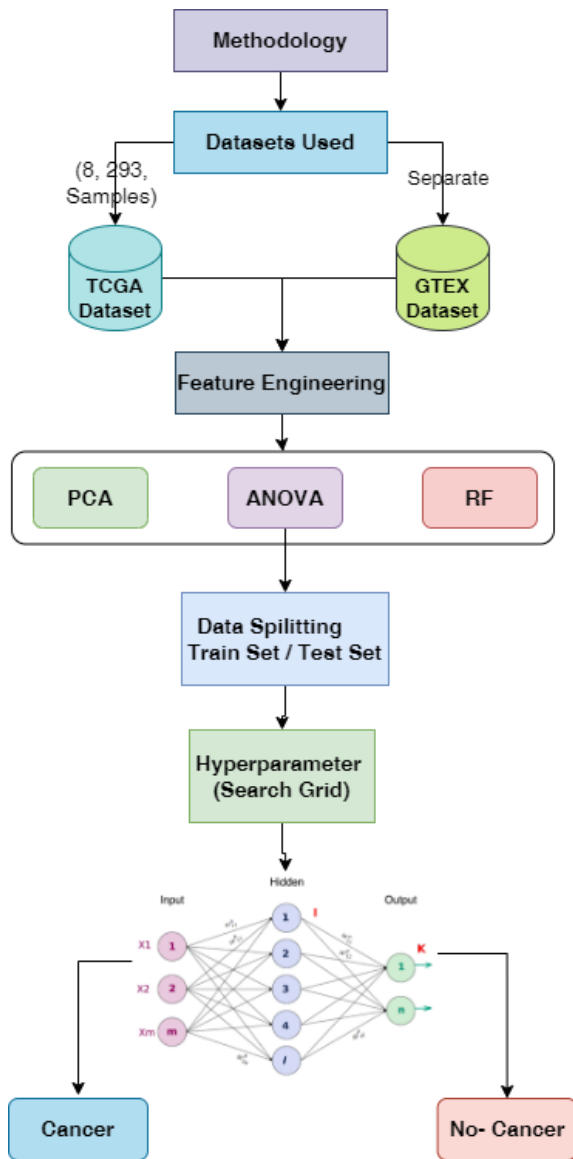


Figure 1. Proposed methodology

3.1 Data sets used

The data used in this study comes from The Cancer Genome Atlas (TCGA), the world's largest repository of genomic data. This collaborative initiative, which was established in 2006, brings together the expertise of the National Cancer Institute's Center for Cancer Genomics and the National Human Genome Research Institute [25], TCGA has significantly contributed to advancements in accurate cancer diagnoses, treatment strategies, and preventive measures. Its public accessibility has greatly facilitated researchers, simplifying their investigations. The TCGA dataset employed herein comprises 8,293 samples

categorized into 15 distinct cancer types. We constructed our model using this dataset and tested it on a separate dataset from the Genotype-Tissue Expression (GTEx) project. The ongoing GTEx project aims to develop a comprehensive public resource for analyzing tissue-specific gene expression and regulatory mechanisms, offering valuable insights into various biological processes [25].

3.1.1 Data description

The dataset obtained from TCGA encompasses 38,019 features and includes 8,293 samples. To facilitate utilization with machine learning models, the data was formatted into a .csv file format. Within this formatted file, the initial column signifies patient IDs, followed by the 'Type' variable representing the target to be predicted. The subsequent columns contain information pertaining to various gene expressions. The 'Type' variable serves as the indicator for cancer type, with each cancer subtype uniquely labeled. The dataset encapsulates 15 distinct cancer types, each labeled in accordance with the TCGA notation, as detailed in Table 1 along with their respective descriptions and labels.

Table 1. Cancer labels and description

Label	Description
STAD	Stomach adenocarcinoma
BRCA	Breast invasive carcinoma
PAAD	Pancreatic adenocarcinoma
ESCA	Esophagus carcinoma
PCPG	Pheochromocytoma and paraganglioma
LUAD	Lung adenocarcinoma
KIRC	Kidney renal papillary cell carcinoma
COAD	Colon adenocarcinoma
UCEC	Uterine corpus endometrial carcinoma
THCA	Thyroid carcinoma
HNSC	Head and neck squamous cell carcinoma
PRAD	Prostate carcinoma
BLCA	Bladder urothelial carcinoma
LIHC	Liver hepatocellular carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma

In this paper, we will focus solely on two types of cancer labels: Breast and Kidney cancers [26].

3.1.2 Feature selection

The process of feature selection plays a pivotal role in model optimization by narrowing down the number of input variables to those most pertinent for predicting target labels. It involves the curation of relevant features from the dataset, significantly influencing the model's performance. In this study, we employed the SelectKBest algorithm from the sklearn library to execute feature selection. This algorithm retains only the top k highest-scoring features, based on a specified scoring function. By utilizing the chi-squared function, pairwise computations between features were conducted. Subsequently, the best features were identified by amalgamating various score values derived from both the Chi-squared test and F-score. This selection process resulted in a reduction of features to 832, which were then utilized to evaluate different models' performance.

3.1.3 Feature extraction

In feature extraction, the primary aim is to identify the most relevant features that carry substantial information to effectively differentiate between distinct classes. While

manual selection of crucial features is considered optimal, in this scenario, it's unfeasible due to the dataset's complexity. Thus, for feature extraction, we employed PCA with varying principal components (200, 300, 500, and 700), ANOVA with selected features (200, 300, 500, and 700). Hence, 606 components were the highest number selected for kidney data, simplifying comparisons across different feature selection methods and ensuring a manageable assessment of the optimal model and feature selection approach.

This paper introduces three prominent feature extraction algorithms commonly employed in genomic classification studies: ANOVA, Random Forest, and PCA. ANOVA (Analysis of Variance), suggested by Bartík [5]. Random Forests, as described in the study conducted by Sathe et al. [16], comprises a set of decision tree classifiers that split datasets based on feature condition values, such as Gini impurity or information gain/entropy, effectively assessing each feature's impact on classification. On the other hand, Principal Component Analysis (PCA) isn't merely a feature selection method but a dimensionality reduction technique. PCA transforms high-dimensional datasets into lower dimensions, enabling faster algorithmic computations and simplified visualization by reducing the dataset's freedom of hypotheses. By focusing on components with the highest variance, PCA allows explaining a significant portion of dataset variance with fewer components, compared to only 89% for the breast dataset. It's crucial to normalize the dataset before applying PCA to ensure each attribute contributes effectively to the analysis.

The selection of ANOVA, Random Forest, and PCA for feature extraction in this study is motivated by their distinct advantages in handling complex and high-dimensional RNA-seq data. ANOVA was chosen for its ability to statistically highlight significant differences between classes, making it a robust method for feature selection in datasets where inter-class variability is crucial. Random Forest was selected for its effectiveness in ranking feature importance and its capability to handle large datasets with many features, providing insights into which features are most influential in classification tasks. PCA was employed to reduce the dimensionality of the dataset while preserving the variance, facilitating more efficient computation and visualization. By focusing on principal components with the highest variance, PCA allows us to manage the high-dimensional nature of RNA-seq data and improve the computational feasibility of subsequent analyses. By combining these methods, we ensure a comprehensive feature extraction approach that leverages the strengths of statistical analysis, machine learning, and dimensionality reduction, ultimately aiming to improve the accuracy and efficiency of cancer subtype prediction.

3.1.4 Data splitting

Using the same dataset for both training and testing can cause overfitting and limit the model's generalization ability. To avoid these issues, the dataset was divided into separate training and test sets. In this study, the TCGA data served for training the models, employing an 80/20 split where 80% of the data was allocated for model training and the remaining 20% for testing. This division ensured that the model learned from a substantial portion of the data while being assessed on unseen data to gauge its performance accurately. Additionally, the independent GTEX dataset, unseen during model training, was employed exclusively for testing purposes, ensuring an unbiased evaluation of model performance. To achieve this

splitting, the sklearn library's train/test split module was utilized, ensuring a systematic and unbiased segregation of the dataset for training and evaluation.

3.1.5 Hyperparameter tuning

Machine learning algorithms often comprise both parameters and hyperparameters. Hyperparameters, unlike parameters, are user-defined settings essential to the model and cannot be learned during training. They remain external and must be pre-determined by the model developer. Preceding the execution of a machine learning algorithm, configuring various hyperparameters becomes imperative. In this study, we employed GridSearchCV, a feature within the sklearn library, to systematically ascertain the optimal hyperparameters for the model. GridSearchCV extensively explores a range of specified parameters through cross-validation, enabling a comprehensive search to identify the best hyperparameters that optimize model performance.

3.2 Deep learning structure

The application of feed-forward neural networks (FFNs) in cancer detection through gene expression data involves constructing a multilayer perceptron with distinct layers representing different aspects of the dataset. The input nodes signify gene expression values, while the output nodes classify samples into cancerous or non-cancerous categories. Through iterative backpropagation during training, the FFN learns intricate patterns within the gene expression data, following preprocessing steps like normalization, feature selection, and dataset partitioning for training and validation. The optimal FFN model consists of three layers: an input layer whose nodes vary depending on the extracted features, a hidden layer with 20 nodes, and an output layer reflecting the two output classes healthy and cancerous. The Rectified Linear Unit (ReLU) serves as the activation function in this model. Figure 2 visualizes the final FFN model structure.

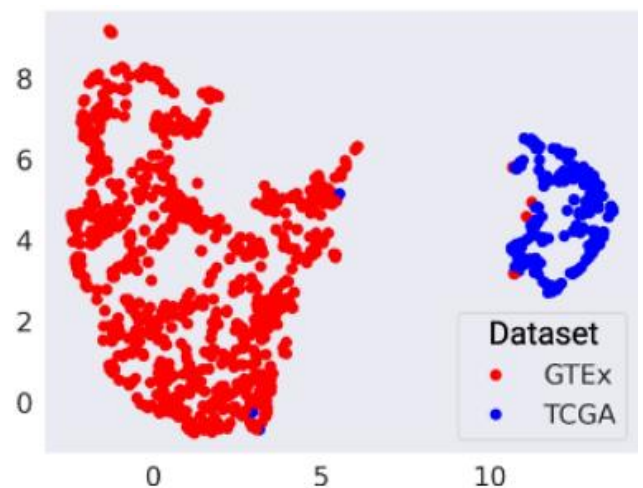


Figure 2. Distribution of cancer labels for TCGA and GTEX

To prevent over-fitting, various combinations of different complexity levels were tested. The best results were achieved using a Feedforward Neural Network (FFN) with three layers: an input layer, a hidden layer, and an output layer. The number of nodes in the input layer varied based on the feature selection method and the number of features extracted. The hidden layer consisted of 20 nodes, while the output layer had 2 nodes,

corresponding to the two classes (healthy and cancerous). The ReLU function was used for activation, with a batch size of 60, trained over 60 to 80 epochs. Stochastic gradient descent was used for optimization, and cross-entropy was used for the loss function. The structure of the FFN model is illustrated in Figure 3, where $I_1...I_n$ indicate the input nodes (n is the number of extracted features), $H_1...H_m$ indicate the hidden nodes ($m=20$), and $O_1...O_2$ indicate the output nodes.

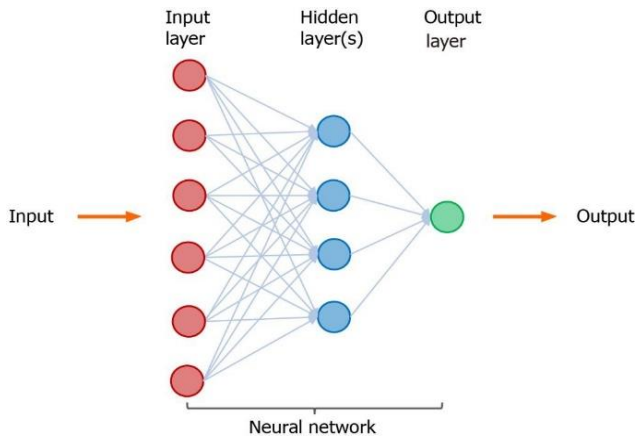


Figure 3. Deep neural architecture

4. EXPERIMENT RESULTS

The TCGA dataset was divided into training and validation sets. The Feedforward Neural Network (FFN) was trained on the training set, and various hyperparameters were tuned using cross-validation. After identifying the optimal hyperparameters, the entire TCGA dataset was used to fit the model. Finally, the independent GTEX dataset was used to test the models. This section will discuss the results from different feature selection algorithms using data representation from the marginalized stacked denoising autoencoder (mSDA).

4.1 Deep neural network results

Figure 4 and Figure 5 display the attained validation accuracy, showcasing the optimal outcomes achieved through distinct feature extraction methods and model structures for TCGA kidney and breast data. The study findings revealed that the use of Principal Component Analysis (PCA) as the feature extraction method, with 300 principal components, in conjunction with a 2-layered neural network structure, yielded the highest accuracy and lowest error rates when analyzing TCGA kidney data. On the other hand, when analyzing TCGA breast data, the most favorable results with the lowest error rates were achieved by employing PCA. These results highlight the importance of tailoring the feature extraction method and neural network architecture to the specific dataset being analyzed. By optimizing these parameters, the study demonstrates the potential to enhance the accuracy and reliability of tumor classification models, providing valuable insights for improving diagnostic and treatment approaches in the context of breast and kidney cancers. These findings underscore the significance of tailoring feature extraction methods and model architectures based on specific dataset characteristics to achieve the most accurate and efficient predictive models.

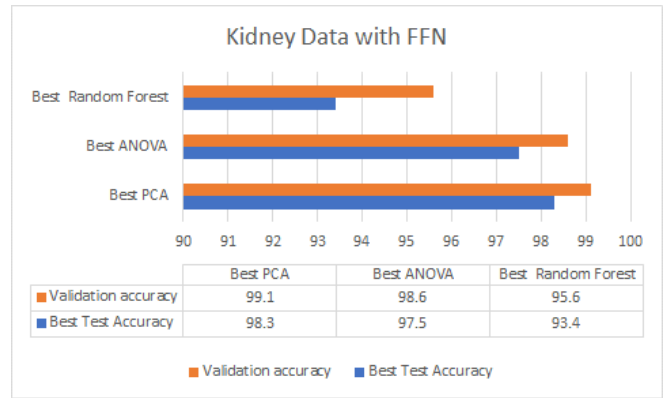


Figure 4. The kidney dataset exhibited the most optimal accuracy

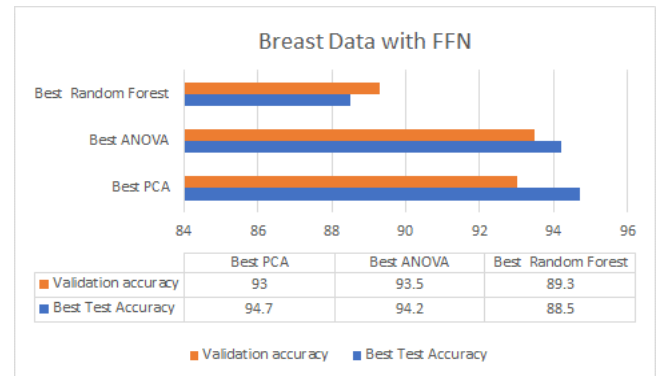


Figure 5. The breast dataset showed the most favorable accuracy

Effective feature extraction methods are crucial for advanced analysis in this context. Deep learning models often confront a scarcity of samples to establish connections among numerous features. The volume of features significantly influences learning accuracy; an excess of features can hinder the model's ability to accurately predict new samples. Figure 5 demonstrates the overarching impact of features on the models' predictive capacity. The optimal number of features for classification tasks varies extensively based on the input data. To elaborate, approximately features are deemed to yield accurate predictions with TCGA breast data, while a reduced set of 300 features suffices for effective predictions with TCGA kidney data. This observation highlights the dataset-dependent nature of feature requirements crucial for the model's predictive performance. The impact of Principal Component Analysis (PCA) on preserving variance is worth mentioning. Notably, within the kidney dataset, it was observed that utilizing 300 principal components is sufficient to retain 94% of the variance. The same number of components only captures 82% of the variance. These findings highlight the inherent differences in gene expression patterns between kidney and breast cancer. The higher percentage of variance retained in the kidney dataset suggests a more concentrated and distinct gene expression profile, whereas the lower percentage in the breast dataset indicates a higher level of heterogeneity. To reach the desired 94% variance in the breast dataset, an additional 400 principal components are essential. This variance discrepancy between the datasets directly impacts the requisite number of features to maintain accuracy, affirming the expectation that the feature count crucial for accurate predictions differs significantly for each dataset.

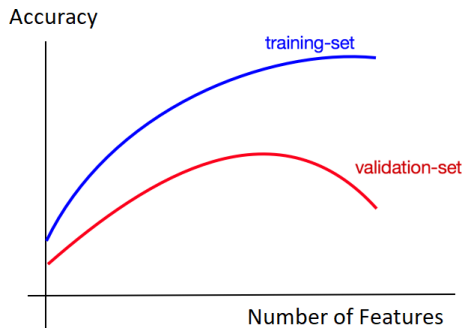


Figure 6. Impact of feature selection on analysis results

The dissimilarities between the kidney and breast datasets also have implications for the complexity of constructing effective models. Interestingly, when training the kidney data, more complex structures tend to lead to overfitting or underfitting issues, necessitating the use of simpler models to achieve satisfactory results. Conversely, with the breast data, it is crucial to employ more complex structures to obtain a decent model. Both datasets are trained for binary classification tasks, but due to their inherent differences, distinct model structures are required to achieve desirable outcomes. In terms of feature extraction methods, PCA demonstrates superior performance, closely followed by ANOVA. Figure 6 displays the impact of feature selection on the analysis results. Notably, the model tends to underfit the healthy samples from the input, which could be attributed to the imbalance in the input data.

4.2 Feature selection variation

The experiments involved varying amounts of labeled data for model training. Initially, when the labeled sample count was too low, for instance, at 2, 10, or 20, the model struggled to improve beyond the null hypothesis. However, once the labeled sample count reached 50, which is relatively higher but still considered small, the model began exhibiting significantly high accuracy.

The supervised learning part of the model received a balanced ratio of labeled data, with 50% representing both cancerous and healthy samples. Interestingly, beyond the threshold of 50 labeled data, there wasn't a noticeable accuracy improvement even with additional labeled samples, such as 200. Surprisingly, the final accuracy between experiments using 50 and 200 labeled data appeared very similar. This suggests that the unsupervised learning aspect functioned effectively alongside the supervised learning, contributing to the model's performance. To ensure consistent training conditions, the batch size for these experiments was set at 60, aligning with the number of labeled data samples fed into the input for trials. This configuration allowed for a seamless and effective integration of both supervised and unsupervised learning components within the model.

In addition to the established choice of 60 labeled data samples from the TCGA kidney and breast datasets, the experiment explored smaller and larger labeled data sample sizes. Employing the same feature selection methods used previously facilitated an easier comparison. The methods included PCA with different counts of principal components (200, 300, 500, 700), ANOVA with varying selected features (200, 300, 500, 700), and random forest with different importance levels (.001, .0005, .0001). The outcomes of this experiment are visually depicted in Figures 7-12.

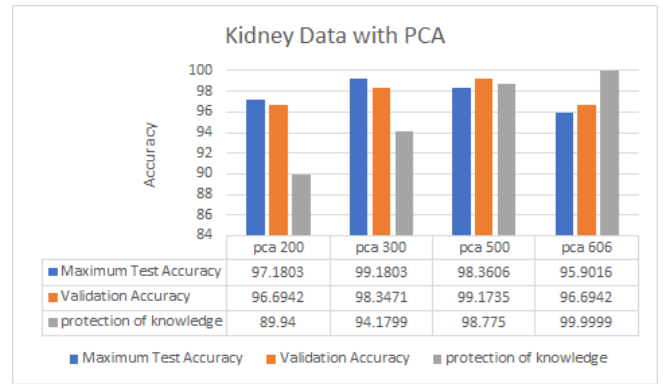


Figure 7. Performance evaluation of feature selection via PCA in kidney data classification

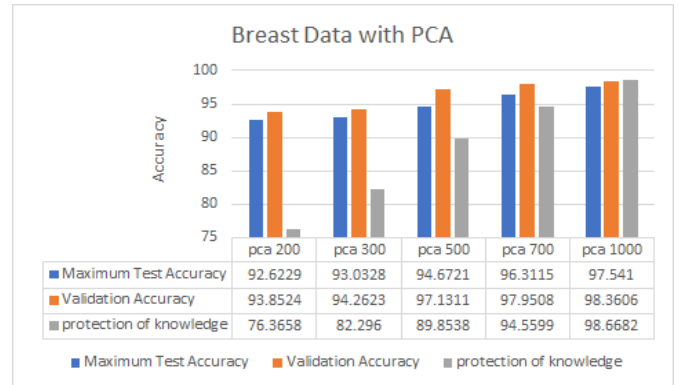


Figure 8. Analyzing accuracy in breast data classification with PCA-based feature selection

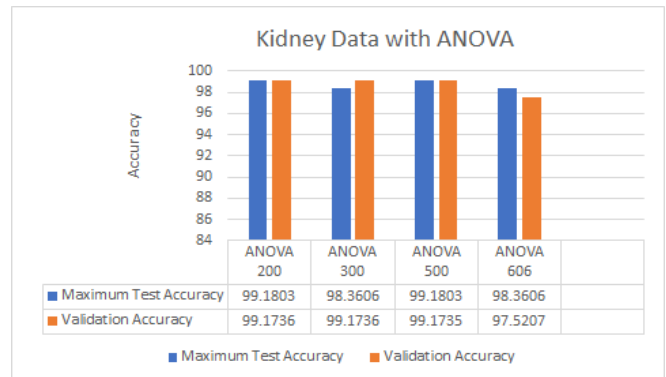


Figure 9. Examining accuracy in kidney data classification with ANOVA-based feature selection

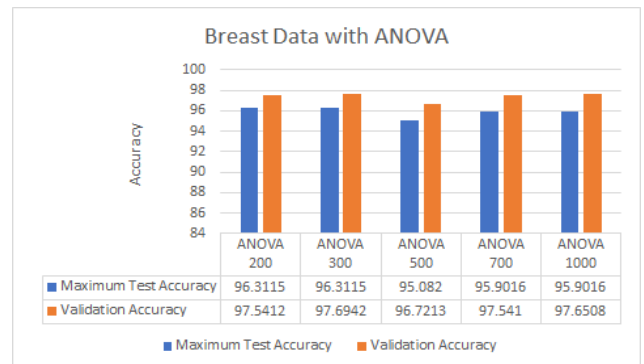


Figure 10. Analysis of accuracy in breast data classification using ANOVA-based feature selection

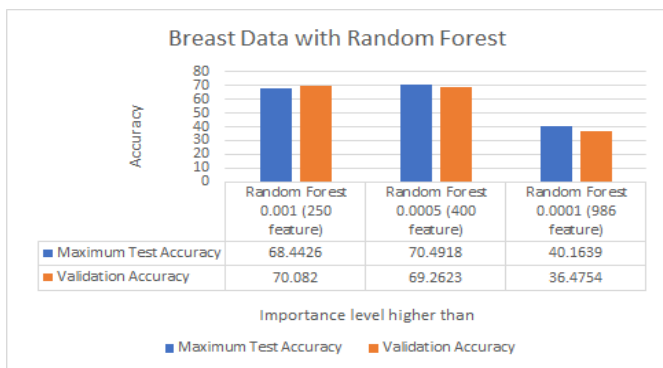


Figure 11. Performance evaluation: kidney data classification accuracy with random forest feature selection

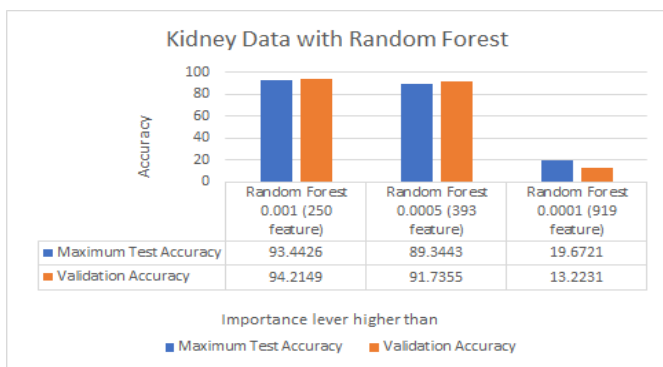


Figure 12. Performance evaluation: breast data classification accuracy with random forest feature selection

The impact of feature extraction aligns closely with the prior findings observed in the FFN experiments. Without applying any feature extraction algorithm, the FFN network struggles to enhance the null hypothesis. However, the FFN network demonstrates stability and achieves better accuracy in most cases. Interestingly, it maintains consistent accuracy regardless of the number of iterations. A detailed analysis, particularly concerning the breast data, accentuates the notable discrepancy between PCA and ANOVA. Specifically, it becomes apparent that ANOVA achieves the highest final accuracy even with just 200 selected features. Conversely, PCA requires a larger count of principal components, ranging from 700 to 1000, to attain a similar accuracy level. Conversely, the features selected using random forest fail to reach even the null hypothesis.

The overall performance of the FFN is shown in Table 2. We achieved 96% accuracy on the TCGA dataset and 80% accuracy on the GTEX dataset. This model produced results similar to KNN but required less computation time. Using the selected features, the model was computationally faster than before. Table 2 presents the classification report of the FFN, indicating ongoing issues with misclassified samples within the GI category.

Table 2. Result of FFN using selectKbest

Model	Multilayer Perceptron
Best hyperparameter	hidden layer sizes=416, activation=tanh, alpha=0.001
Best training accuracy	98.36%
Test set accuracy (TCGA)	98.66%
Accuracy on independent data (GTEX)	97.54%

4.3 Comparison

In this study, a feed-forward network model was utilized to perform binary classification on TCGA breast and kidney data. The aim was to attain the highest possible accuracy by exploring and comparing different feature selection methods and their respective feature quantities. By employing these methods, the researchers sought to identify the most effective combination of features that would yield optimal results. Additionally, the study aimed to determine which feature selection methods were most suitable for each dataset. This comprehensive analysis allowed for a thorough evaluation of the performance and accuracy of the feed-forward network model in the classification of breast and kidney data from the TCGA dataset. The raw TCGA kidney dataset contains 606 samples and 20,206 genes, while the TCGA breast dataset consists of 1,218 samples and 20,207 genes. The data was divided into training (60%), test (20%), and validation (20%) sets, resulting in 363 samples for kidney data training, and 730 samples for breast data training. The testing and validation sets comprised 122 and 244 samples for kidney and breast data, respectively. To address data imbalance issues, the input data underwent batch processing following the same ratio across both used structures.

The feed-forward network emerges as a favorable choice, especially for those less acquainted with deep learning, offering a clear and modifiable structure that allows easy experimentation. While it attained an acceptable accuracy rate with the kidney data, attempts to stabilize the results faced challenges, mainly stemming from imbalanced input data, leading to frequent biased predictions favoring cancerous data.

The FFN's instability could be due to several factors such as an inappropriate learning rate, batch size, and weight initialization, as well as overfitting and the inherent variability in Stochastic Gradient Descent (SGD). To mitigate these issues, implementing a learning rate scheduler or using adaptive learning rate methods like Adam, adjusting the batch size, employing advanced weight initialization techniques, and incorporating regularization methods like dropout and L2 regularization are recommended.

The impact of different feature extraction algorithms varied across the selected structures and datasets. With the kidney data in the FFN structure, the principal components selected by PCA delivered the highest accuracy, closely followed by ANOVA, whereas Random Forest yielded the poorest results, reaching only 95% accuracy. Conversely, with the breast data, Random Forests showcased the best performance, Random Forest's efficacy differed significantly, implying its superior performance with more intricate datasets, such as the breast dataset. To maintain 94% variance preservation in the breast dataset, an additional 400 principal components were necessary. Enhancements in accuracy may ensue from employing diverse feature extraction methods or manual selection by domain experts. Future research avenues could explore multi-label classification as an alternative to binary classification

5. CONCLUSION

In biological data analysis, machine learning methods, especially deep learning, have gained attention. Deep learning's performance is notable; however, its reliance on substantial data can limit accuracy in data-limited scenarios.

This paper introduces FFN network structures for binary classification using TCGA mRNA data for breast and kidney cancer. The FFN achieves 99.2% accuracy with the kidney dataset, showing promising outcomes. Future work may expand this study to include various TCGA datasets, enabling experiment with different deep learning architectures beyond FFN, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to capture complex patterns in genomic data more effectively. These architectures may offer better performance and stability for cancer classification tasks, further refine the feature selection process by exploring additional methods or combinations of methods, investigate the impact of feature selection techniques on model stability and performance across different cancer types and datasets.

REFERENCES

- [1] Munkácsy, G., Santarpia, L., Györfy, B. (2022). Gene expression profiling in early breast cancer—Patient stratification based on molecular and tumor microenvironment features. *Biomedicines*, 10(2): 248. <https://doi.org/10.3390/biomedicines10020248>
- [2] Awaludin, A.M., Larasati, H.T., Kim, H. (2021). High-speed and unified ECC processor for generic Weierstrass curves over GF (p) on FPGA. *Sensors*, 21(4): 1451. <https://doi.org/10.3390/s21041451>
- [3] Bhandari, N., Walambe, R., Kotecha, K., Khare, S.P. (2022). A comprehensive survey on computational learning methods for analysis of gene expression data. *Frontiers in Molecular Biosciences*, 9: 907150. <https://doi.org/10.3389/fmolb.2022.907150>
- [4] Khalsan, M., Machado, L.R., Al-Shamery, E.S., Ajit, S., Anthony, K., Mu, M., Agyeman, M.O. (2022). A survey of machine learning approaches applied to gene expression analysis for cancer prediction. *IEEE Access*, 10: 27522-27534. <https://doi.org/10.1109/ACCESS.2022.3146312>
- [5] Bartik, M. (2020). External power gating technique—An inappropriate solution for low power devices. In 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, Canada, pp. 0241-0245. <https://doi.org/10.1109/IEMCON51383.2020.9284855>
- [6] Demkow, U., Ploski, R. (2015). *Clinical Applications for Next-Generation Sequencing*. Academic Press.
- [7] Kumar, A., Halder, A. (2020). Ensemble-based active learning using fuzzy-rough approach for cancer sample classification. *Engineering Applications of Artificial Intelligence*, 91: 103591. <https://doi.org/10.1016/j.engappai.2020.103591>
- [8] Ban, M., Petrić Miše, B., Vrdoljak, E. (2020). Early HER2-positive breast cancer: Current treatment and novel approaches. *Breast Care*, 15(6): 560-569. <https://doi.org/10.1159/000511883>
- [9] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3): 209-249. <https://doi.org/10.3322/caac.21660>
- [10] Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., Colliot, O., Alzheimer's Disease Neuroimaging Initiative, the Australian Imaging Biomarkers and Lifestyle Flagship Study of Ageing. (2020). Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, 63: 101694. <https://doi.org/10.1016/j.media.2020.101694>
- [11] Elbashir, M.K., Ezz, M., Mohammed, M., Saloum, S.S. (2019). Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression data. *IEEE Access*, 7: 185338-185348. <https://doi.org/10.1109/ACCESS.2019.2960722>
- [12] Monti, M., Fiorentino, J., Milanetti, E., Gosti, G., Tartaglia, G.G. (2022). Prediction of time series gene expression and structural analysis of gene regulatory networks using recurrent neural networks. *Entropy*, 24(2): 141. <https://doi.org/10.3390/e24020141>
- [13] Bar-Joseph, Z., Gitter, A., Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13(8): 552-564. <https://doi.org/10.1038/nrg3244>
- [14] Lee, H.J., Chung, Y., Chung, K.Y., Kim, Y.K., Lee, J.H., Koh, Y.J., Lee, S.H. (2022). Use of a graph neural network to the weighted gene co-expression network analysis of Korean native cattle. *Scientific Reports*, 12(1): 9854. <https://doi.org/10.1038/s41598-022-13796-9>
- [15] Zhang, T.H., Hasib, M.M., Chiu, Y.C., Han, Z.F., Jin, Y.F., Flores, M., Chen, Y., Huang, Y. (2022). Transformer for gene expression modeling (T-GEM): An interpretable deep learning model for gene expression-based phenotype predictions. *Cancers*, 14(19): 4763. <https://doi.org/10.3390/cancers14194763>
- [16] Sathe, S., Aggarwal, S., Tang, J. (2019). Gene expression and protein function: A survey of deep learning methods. *ACM SIGKDD Explorations Newsletter*, 21(2): 23-38. <https://doi.org/10.1145/3373464.3373471>
- [17] Gunavathi, C., Sivasubramanian, K., Keerthika, P., Paramasivam, C. (2021). A review on convolutional neural network based deep learning methods in gene expression data for disease diagnosis. *Materials Today: Proceedings*, 45: 2282-2285. <https://doi.org/10.1016/j.matpr.2020.10.263>
- [18] Tabares-Soto, R., Orozco-Arias, S., Romero-Cano, V., Bucheli, V.S., Rodríguez-Sotelo, J.L., Jiménez-Varón, C.F. (2020). A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Computer Science*, 6: e270. <https://doi.org/10.7717/peerj-cs.270>
- [19] Li, F., Cheng, D., Liu, M. (2017). Alzheimer's disease classification based on combination of multi-model convolutional networks. In 2017 IEEE International Conference on Imaging Systems and Techniques (IST), Beijing, China, pp. 1-5. <https://doi.org/10.1109/IST.2017.8261566>
- [20] Mazlan, A.U., Sahabudin, N.A., Remli, M.A., Ismail, N.S.N., Mohamad, M.S., Nies, H.W., Abd Warif, N.B. (2021). A review on recent progress in machine learning and deep learning methods for cancer classification on gene expression data. *Processes*, 9(8): 1466. <https://doi.org/10.3390/pr9081466>
- [21] Thakur, T., Batra, I., Luthra, M., Vimal, S., Dhiman, G., Malik, A., Shabaz, M. (2021). Gene expression-Assisted cancer prediction techniques. *Journal of Healthcare*

- Engineering, 2021: 4242646.
<https://doi.org/10.1155/2021/4242646>
- [22] Das, S., Rai, A., Merchant, M.L., Cave, M.C., Rai, S.N. (2021). A comprehensive survey of statistical approaches for differential expression analysis in single-cell RNA sequencing studies. *Genes*, 12(12): 1947. <https://doi.org/10.3390/genes12121947>
- [23] Li, S., Xu, X., Zhang, R., Huang, Y. (2022). Identification of co-expression hub genes for ferroptosis in kidney renal clear cell carcinoma based on weighted gene co-expression network analysis and the cancer genome atlas clinical data. *Scientific Reports*, 12(1): 4821. <https://doi.org/10.1038/s41598-022-08950-2>
- [24] Divate, M., Tyagi, A., Richard, D.J., Prasad, P.A., Gowda, H., Nagaraj, S.H. (2022). Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures. *Cancers*, 14(5): 1185. <https://doi.org/10.3390/cancers14051185>
- [25] Vaiyapuri, T., Liyakathunisa, Alaskar, H., Aljohani, E., Shridevi, S., Hussain, A. (2022). Red fox optimizer with data-science-enabled microarray gene expression classification model. *Applied Sciences*, 12(9): 4172. <https://doi.org/10.3390/app12094172>
- [26] The Cancer Genome Atlas (TCGA). National Cancer Institute. <https://www.cancer.gov/toga>, accessed on 13 Sep. 2023.