Feedforward Deep Learning Optimizer-based RNA-Seq Women's Cancers Detection with a Hybrid Classification Models for Biomarker Discovery

Waleed Mahmoud Ead[1], Marwa Abouelkhir Abdelazim[2], Mona Mohamed Nasr[3]
Information Systems Department-Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt[1, 2]
Information Systems Department-Faculty of Computers and Artificial Intelligence, Helwan University, Egypt[3]

***Abstract*—Women's cancers, signified by breast adenocarcinoma and non-small-cell lung cancers, are a significant threat to women's health. Across the globe, the leading cause of death for women is a group of tumors referred to as "female-oriented cancers". The most recent researches in the classification of molecular tumors is the analysis of women's cancers using RNA-Seq data for precision cancer diagnoses. Furthermore, discovering the different genes' patterns behaviors will lead to predict the cancer-specific biomarkers to early diagnosis and detection of cancer-specific in women. An overfit model will be resulted due to the high-dimensional data of RNA-Seq from a small samples of data. In this work, we propose a filter-based selection approach for a deep learning-based classification model. In addition, hybrid classification models have been proposed to be compared with the new modified deep learning model. The Experiments' analysis showed that the proposed filter-based selection approach for a deep learning-based classification model performed better than the other hybrid models in terms of performance evaluation metrics, with an accuracy of 96.7% for RNA-Seq breast adenocarcinoma data and 95.5% for RNA-Seq non-small-cell lung cancer data.**

Keywords—Women's cancers; RNA-Seq; deep learning; molecular tumor; hybrid classification models

I. Introduction

Today, cancer is the number two mortality globally and the number one mortality in both developed and developing nations. Twenty million new cases of cancer are diagnosed each year, and ten million people die from it. Women account for nine million of these cases and 4.4 million deaths globally each year [1]. According to the World Health Organization, breast cancer will be the most prevalent and deadly cancer in women in 2020, accounting for more than two million new cases and 684,996 deaths annually. Lung cancer is the third most frequent type of cancer in women overall, with roughly 607, 465 deaths and 770, 828 cases every year [1].

Determining the presence of cancer, making a primary diagnosis, and identifying new, more effective treatment options could all aid in reducing mortality and morbidity rates. Genetics, individual lifestyle, body shape, age, menopause status, family history, smoking, and history of exposure to carcinogens or viruses, hormone therapy, chemicals, and other airborne particles are all associated with the occurrence and frequency of women's cancers [2]. One of the most crucial methods to investigate genetic correlations in medical investigations is transcriptomics by next-generation RNA sequencing (RNA-seq). Large data sets generated by NGS technologies offer a thorough perspective of the human genome [3]. Numerous molecular structures are adopted by nucleic acids, and these designs are crucial for the storage, processing, and transmission of genetic information [4]. DNA molecules are translated to mRNA for the synthesis of proteins. Proteins are the primary factors in the most fundamental cellular processes. The process via which a fragment of DNA is read and then transformed into an addiction to a protein has excessive awareness in several therapeutic analyses in addition to biological ones [5]. The significant aim of cancer disease research is to recognize the genes that cause normal cells to mutate into cancer [6].

Researchers now have access to an unprecedented amount of tumor genomic and transcriptome data thanks to developments in next-generation sequencing methods. A molecule that is reliably tested and assessed as a marker of healthy biological processes, harmful biological processes, or pharmacologic reactions to therapeutic intervention is referred to as a biomarker [7]. Next-generation sequencing (NGS) technologies are used by RNA-Seq. As a crucial tool, RNA sequencing has been used in many aspects of cancer research and therapy, including the identification of biomarkers and the characterization of cancer heterogeneity and evolution, drug resistance, the cancer immune microenvironment, and immunotherapy, among others [8].

It is critical to develop biomarkers for disease progression and potential therapy response, as this will enable personalized care, enhance clinical outcomes, and accomplish the objective of precision oncology [9]. This kind of technology is gathering data from cells and tissues about variations in gene expression [10]. Depending on whether targeted-exome or whole-exome sequencing is utilized, the potential of RNA sequencing resides in the ability to combine the twin characteristics of discovery and quantification in a single high-throughput sequencing, allowing for the simultaneous investigation of thousands of genes [11]. Finding the set of genes that are associated with and highly expressed in many types of tumor cells is one of the difficult issues in the field of cancer classification [12]. Massive gene data sets with few samples are frequently used to represent gene expression data [13]. The large number of ambiguous and redundant features in gene data has been highlighted as adding to the classifiers' complexity challenges.

There are two significant problems with the RNA-Seq gene expression datasets [13]. Due to the high dimensionality of the RNA-Seq datasets, the datasets are extremely complicated and noisy [14]. In these datasets, only a small number of samples were gathered, even though each sample measures the levels of expression of countless thousands of genes [14]. As a result, the learning model will be overfit due to the dimensionality problem and the problems with such a large dataset. Using gene expression analysis, researchers may categorise malignancies, forecast clinical outcomes, and identify biomarkers connected to the disease. The current main hurdle in the cancer diagnosis problem is thought to be the differentiation of normal from malignant tissues, as well as the selection of the few informative genes [15].

In this work proposed RNA-Seq gene expression classification models that are optimised for deep learning and combined with PDA, SVMRadial, GaussprPoly, NB, RF, NN, and the Glmboost Method. The NCBI GEO accessions GSE19804 and GSE70947 were used to download the RNA-Seq gene expression profiles. And extensive packages that make RNA-Seq analysis possible when using Bioconductor and R programming. It has five modules: feature mapper, preprocessing gene expression, dimension transformer, feature selector, deep learning approach, machine predictors with hyper-parameters, and prediction biomarkers with performance evaluation including accuracy, sensitivity, specificity, precision, the F1 score, and the area under the curve (AUC) score.

The contribution of this work can be summarized as follows:

- Using supervised learning and a deep learning algorithm known as a feedforward neural network with hyper-parameters for model optimization,
- For the selection approach, we introduce filter-based selection for dimensionality reduction methods for selection informative genes by applying the FCBF algorithm.
- To determine the dependencies of genes and identify the optimal subset of genes, the enhanced gene selector and feed-forward neural network classifier combined the statistical results of pertinent genes using the symmetrical uncertainty (SU) assessment.
- We adopted further classification. A model that achieves robust classification with little CPU consumption while maintaining accuracy under test conditions is based on hybrid learning models with feedforward neural networks.

The rest of this paper is organized as follows. Section II discussed some of related works, Section III discussed the materials and methods used in this work, and Section IV illustrates the proposed approach, while Section V discusses our results. Finally, Section VI provides conclusion.

II. Related Work

Different related works have been proposed in the era of detection and the diagnosis of human cancer, and in this section reviews the most recent studies on the use of deep learning and machine learning in the field of malignant tumor gene expression data. Studies on biomarker gene documentation will also be tested. Researchers will be able to evaluate and appraise their suggested analytical methodologies using data on cancer gene expression from the resources they have identified.

Zhang et al. [16], proposed a SVM classifiers based on various features selection to forecast lymphatic metastasis in a range of malignancies. Such classifiers were applied to identify differentially expressed signatures in lymph node metastatic and non-metastatic cancer groups. These SVM classifiers were found to be successful, with an overall accuracy of 81.25% on various profiles with light biomarker sets (seven biomarkers on average). They also contrasted these SVM classifiers with two other benchmark classifiers based on comparable profiles (Random Forest, KNN, and K-Nearest Neighbor, RF).

Han et al. [17], argue that Rao's score statistic is arithmetically appropriate to associate several mechanisms through a typical set of weight factors to yield a biased universal indicator. Next, the weightiness slash statistics measure the purposeful influences of various alteration categories on the target population. Finding cancer-associated genes with mutations that cause the cancer phenotype during cancer genome sequencing is a significant issue for this paper.

Simsek et al. [18], proposed the machine learning classification model for the classification of leukemia subtypes using the gene expression data set from 72 patient records and 7129 gene regions. In the research, machine learning classification techniques such as support vector machines (SVM), linear discriminant analysis (LD), ensemble classifiers (EC), and K-nearest neighbor (KNN) were used. It is evident from experience that the SVM model outperforms the other algorithms despite the fact that the collective test data and collective training data were shared and combined to create a fresh training dataset. Results show that these machine learning models can be helpful in determining the leukemia subtype.

Das et al.'s [19], grouping and classification approach by gene-gene similarity matrix is permeated by the suggested study. The feature selection strategy in this attempt is SVM-RFE. Based on the gene-gene similarity matrix, specific traits are further clustered into several groupings. These pairwise correlation-based clusters use reduction into a smaller set of features for further processing by the neural network, which is required to categorise the various types of cancer.

Yin et al. [20], The CNN-Cox model, which combines a unique CNN framework with prognosis-related feature selection cascaded Wx and has the advantage of fewer computing requests while using light training parameters, has been established as a short and effective survival analysis model. The Cancer Genome Atlas cohort's seven cancer type datasets, including those for head and neck squamous cell carcinoma, bladder carcinoma, brain low-grade glioma, kidney renal cell carcinoma, skin cutaneous melanoma, lung squamous cell carcinoma, and lung adenocarcinoma (LUAD), show that the CNN-Cox model achieved reliable higher C-index values and better survival prediction performance. They demonstrated the use of protein-protein interaction network analysis to identify potential prognostic genes and further investigated the biological roles of 13 core genes, whose high expression is significantly associated with poor survival in LUAD patients.

Houssein et al. [21], proposed an algorithm to choose the most relevant and instructive genes from cancer microarray datasets. The first goal of this study is to select the most predictive genes, and the second goal is to extract the most accurate gene expression datasets with the least amount of difficulty. The most informative genes are chosen from a tiny, filtered dataset that is collected from the IG feature subset evaluator after filtering out irrelevant and noisy genes and getting their relationships from the datasets using the BMO method with the SVM classifier. To calculate this suggested model's efficiency, four benchmark microarray datasets—namely, Leukemia1, Leukemia2, Lymphoma, and SRBCT—were used.

Vaiyapuri et al. [22], proposed a new Red Fox optimizer for deep learning-supported microarray gene expression classification (RFODL-MGEC).The current RFODL-MGEC model aims to improve classification performance by selecting appropriate features. The RFODL-MGEC model employs a novel feature selection technique based on the red fox optimizer (RFO) with the goal of creating an ideal subset of characteristics. A bidirectional cascaded deep neural network (BCDNN) created for data classification is also part of the RFODL-MGEC model.

Shen et al. [23], report that DCGN, a deep learning method, has been proposed for cancer multi-classification tasks; this model is recommended since it can handle high-dimensional sparse gene expression data better than previous models that have been put out. The DCGN performs well on all five of the examined datasets when it comes to classification evaluation factors like accuracy and precision, especially on the BLCA-TCGA and BLCA-CIT datasets.

Rezaee et al. [24], proposed a hybrid method that assigns rank to the five key genes in the microarray data based on soft ensemble and stacking auto-encoders. The least number of genes needed for final classification was found by combining the three soft wrapper techniques with classification using the k-NN algorithm.

At the conclusion of this section, In this work, we are interested in diseases that affect women, and therefore we strive to provide the latest technologies that help in the early detection of these diseases in order to speed up the treatment process and help doctors take accurate measurements and develop medicines suitable for each disease as Targeted therapies are determined according to the biomarkers for women's carcinoma that were discovered through powerful learning models suitable to deal with high-dimension RNA-Seq gene expression with hyper-parameters to optimizing the model, so this model in our study gave the best results with performance evaluation of classification models on test datasets for women's cancer.

III. Materials and Methods Background

A. *Women's Cancers RNA-Seq Gene Expression Datasets*

Biotechnology National Center Information is a key resource for multi-omics research, including genetic data, and it facilitates the advancement of science and health by making biomedical and genomic information accessible. For example, genome, transcriptome, epigenetic, and proteome information are applied to methodology issues in bioinformatics. Fresh genetic structures (i.e., RNA, DNA, ChIP sequence, whole exome sequencing, protein chips, and amino acid structures) are among the most abundant public raw data in omics and are easily accessible via the following group sequencing tools. The RNA-Seq gene expression profiles used in our investigation were downloaded from the Gene Expression Omnibus (GEO) database, a free public database that included various genes (https://www.ncbi.nlm.nih.gov/geo/). Under the accession numbers GSE19804 and GSE70947, the dataset was downloaded. The explanation for each kind of tumor is given in Table I.

B. *Methods and Materials*

In this section, we examine the various methodologies used for the proposed model.

1) Feature mapping: For each row in the gene expression dataset, feature mapping translates the Entrez Gene id to the gene symbol and gene name before using the merging procedure to connect those annotations [25].

2) Preprocessing: Combinations of normalization-transformation and the imputation method. Normalization is a crucial step in the interpretation of RNA-Seq data since normalization- transformation combinations are regulated by preprocessing. To enable samples to be evaluated on the same scale, systematic deviations must be identified and corrected [26]. These systematic changes may result from both within-sample variations such as gene length and sequence composition as well as between-sample variations such as library size (sequencing depth) and the presence of majority fragments. Additionally, for data compatibility, transformations are used. In order to handle missing values in gene expression, imputation is utilized.

TABLE I. WOMEN'S CANCERS RNA-SEQ GENE EXPRESSION DATASETS

Accession number	Dataset name	Number of features (genes)	Number of samples	Summary
GSE19804	non-small-cell lung cancer (NSCLC)	54,675 features	120 samples	Even though smoking is the main risk factor for lung cancer, in Taiwan, just 7% of female lung cancer patients had ever smoked, a significantly lower percentage than among Caucasian females. This study provides a thorough examination of the molecular profile of female lung cancer in Taiwan that is not caused by smoking.
GSE70947	breast adenocarcinoma	62,976 features	296 samples	Through accelerating angiogenesis and tissue remodeling in the tumor microenvironment, chronic inflammation aids in the growth and invasion of breast tumors. The intricate interaction between estrogen, which promotes the growth of 70% of breast cancers, and inflammation.

3) The dimension transformer. After the RNA-Seq reads have been mapped to a reference genome or transcriptome, the number of reads mapped to the reference genome can be tallied to determine the abundance of the transcripts. For the approaches to be used, it is crucial that the count values be raw sequencing read counts [27].

4) Feature selector: The high dimensionality of the dataset is one of the main issues with machine learning [24]. The weighting features reduce processing time and redundant data, boosting algorithm performance because the analysis of several features uses a lot of memory and results in overfitting [20]. The method of eliminating all unnecessary and irrelevant genes while also identifying the most informative genes [28]. Finding the group of genes that are associated and highly expressed in many types of tumor cells is one of the other difficult issues in the field of cancer categorization [29]. Gene expression data is frequently characterized by an enormous amount of gene data and a small number of samples. It has become clear that the abundance of confusing and duplicated features restricted in the gene data adds to the classifiers' difficulty. To improve the accuracy of predictive models, this research proposed FCBF filter-based dimensionality reduction approach as a selection method. Gene prioritization, often known as the finding of biomarkers, is another name for the feature selection method.

5) Supervised methods: The cross-validation concept: A method for minimizing bias in the estimation of prediction accuracy is cross-validation [6]. When a classification system is over fitted to a certain dataset, bias might result because the algorithm learns the classification "by heart" but struggles to generalize it to new, untested samples. In a nutshell, the dataset is deterministically divided into a number of training and test sets for cross-validation [18]. Each training set is used to build the model, which is then tested on the test set. Over these fits, the accuracy metrics are averaged. N fits are used in leave-one-out cross-validation, with N training sets of size N-1 and N test sets of size 1. As a result, we employed 10-fold cross-validation to define training control in this study's two data sets, splitting the data randomly into a test set (30% of the dataset) and a train set (70% of the dataset).

a) Learning models: Classification and regression are two examples of supervised learning tasks that attempt to anticipate the intended output based on the input data [30]. For instance, a classification algorithm trained on a dataset of correctly classified genes using supervised learning will learn to recognize diseases. In this study, supervised machine learning algorithms with tuning parameters for the gene expression of women's tumors were incorporated in seven learning models to predict biomarkers.

- Neural Network (NN).
- Support Vector Machines with Radial Basis Function Kernel (SVMRadial).
- Penalized Discriminant Analysis (PDA).
- Naive Bayes (NB).
- Random Forest (RF).
- Gaussian Process with Polynomial Kernel (GaussprPoly).
- Boosted Generalized Linear Model (Glmboost).

b) Feedforward neural network algorithm (FNN): An input, multi-hidden, and output hierarchy shape is the Multilayer Perceptron Architecture's most noticeable feature at first glance (Fig. 1). If input data is provided, the output result is computed directly along the subsequent layers of a multilayer perceptron. This type of neural network operating process is referred to as feedforward [31]. A number is obtained as the current output of each neuron in the middle-hidden layer by multiplying the vector-format output results from the previous layer by a weight vector plus a bias value in the current layer, then feeding the biased weighted sum into a nonlinear function (such as a sigmoid, hyperbolic tangent, or rectified linear unit (ReLU), etc.). The feature layer is a new numeric vector made up of enormous neuron outputs in the same hidden layer.

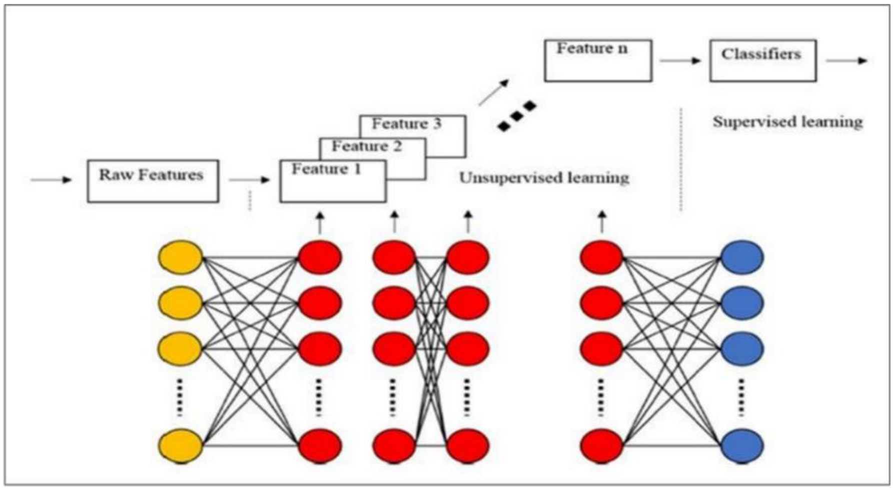

Fig. 1. Multilayer perceptrons architecture.

Evaluation Metrics for Classification Models on test dataset: Several metrics are used to evaluate machine learning approaches. The optimal models are designated using these metrics [30]. To systematically determine the detection effect, the metrics are often used concurrently in our proposed approach as follows:

- Accuracy: A test's accuracy is determined by how well it can distinguish between cancer and healthy instances. Accuracy = TP+TN/TP+TN+FP+FN.
- Sensitivity: A test's sensitivity is how well it can identify cancer instances (true positive rate). Sensitivity = TP / TP + FN.
- Specificity: A test's specificity is how well it can identify healthy instances (false positive rate). Specificity = TN / TN + FP.
- Precision is the ratio of the number of true positive findings to the number of positive results the classifier anticipated. Precision equals TP / TP + FP.
- The F1 score is a direct reflection of the model's performance and is used to evaluate test accuracy. The F1 score can vary from 0 to 1, and the objective is to reach as near to 1 as possible.
- Receiver operating characteristic curve (ROC) / area under curve (AUC) score: The performance of the classification model at every threshold is shown graphically by the ROC curve. The entire region below the ROC curve in two dimensions is known as the AUC. Sensitivity and specificity, two crucial parameters, are produced by this curve.

IV. Proposed Approach

As shown in Fig 2, we employed the RNA-Seq features as the inputs for deep learning-based classification together with other well-known techniques such as PDA, SVMRadial, GaussprPoly, NB, RF, and glmboost to predict biomarkers for women's malignancies.

A. Proposed Approach for Women's Cancers Classification

1) Feature mapper module: By using an organism-level package (an "org" package) that employs a central gene identification (such as the Entrez Gene id) and provides mappings between this identifier and other types of identifiers, annotations can be provided in packages curated by Bioconductor (e.g., GenBank or Uniport accession number, etc.). The number of reads mapped to the reference genome can be tallied to determine the abundance of the transcripts once the RNASeq reads have been mapped to a reference genome or transcriptome. For the approaches to be used, it is crucial that the count values be raw sequencing read counts.

2) A preprocessing module: The normalization-transformation combinations are controlled by using transformations for data compatibility for two reasons: Making non-numeric features into numeric features. Since a string cannot be multiplied using a matrix, it must be converted to a representation that is practically numerical. Resizing inputs to a consistent size. For instance, feed-forward neural networks require input data to be a constant size since they have a certain number of input nodes.

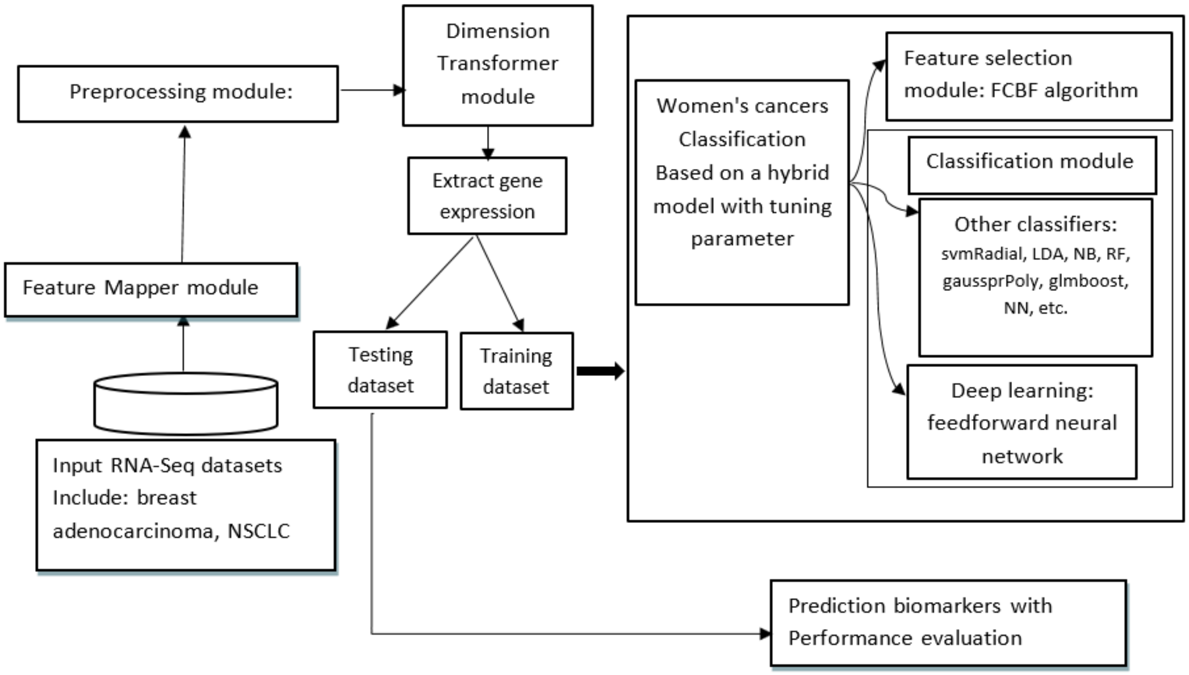
Preprocessing module:
Dimension Transformer module
Extract gene expression
Testing dataset
Training dataset
Feature Mapper module
Input RNA-Seq datasets Include: breast adenocarcinoma, NSCLC
Women's cancers Classification Based on a hybrid model with tuning parameter
Feature selection module: FCBF algorithm
Classification module
Other classifiers: svmRadial, LDA, NB, RF, gaussprPoly, glmboost, NN, etc.
Deep learning: feedforward neural network
Prediction biomarkers with Performance evaluation

Fig. 2. The pipeline of RNA-Seq female cancer data-based machine learning and deep learning workflow development

NAs in this data collection are another problem. That's typical. I'll explain. Without showing log transformations, the expression data have a very wide range, with several outliers in the higher range. However, as log-transformation frequently produces data with negative infinity (-Inf) values or NAs, which are brought on by all the 0-values in the data since log2 (0) == -Inf, this frequently creates a new issue. As a result, set their NAS value to zero.

Methods used for normalization and transformation: deseq-rlog available in MLSeq package in R programming: Deseq median ratio approach is used for normalization. The normalized data is transformed using a regularized logarithmic formula.

z-Score Normalization (zero-mean Normalization)

3) FCBF-PCA Feature selector module: subsequently applied PCA using the FCBF-reduced datasets. While feature selection shrinks the dataset by deleting useless characteristics, dimensionality reduction uses feature extraction to reshape and simplify the data. Use Bioconductor's FCBF (Fast Correlation Based Filter for Feature Selection) to filter highly correlated genes. A multivariate gene selection technique called a fast correlation-based filter (FCBF) begins with a complete collection of characteristics (genes). It determines the optimum subset by calculating the dependencies of the genes using the symmetrical uncertainty (SU) measurement. An effective computer approach called FCBF is used to discriminate between redundant and irrelevant features.

It assesses each property individually, finds the main correlations, and heuristically eliminates superfluous features. When there are no features, it stops due to an internal halting requirement. Implementing FCBF for GSE70947 and GSE19804 datasets seems like 0.1 as a threshold that is reasonable for this both two datasets. After running FCBF for GSE70947, we went from 62976 features to now a lean set of twenty-four features/genes. As can see, EZH2.1 has the strongest correlation to the target class with an SU value of 0.41, and then comes COL10A1.1 with an SU value of 0.40, and so on. As demonstrated, Table II shows the best informative genes and FCBF for GSE19804, we went from 54,675 features to now a lean set of seventy-nine features/genes. As can see, COL10A1 has the strongest correlation to the target class with an SU value of 0.72, and then comes PROM2 with an SU value of 0.68, and so on. As demonstrated, Table III shows the best informative genes

Subsequently, run the FCBF algorithm using a heat map plot to illustrate the gene correlation of the 79 genes for GSE19804 gene expression: Observing the heat map in Fig. 3, we can see the genes are not either positively or negatively correlated with each other as they appear in a lighter color (blue = negative correlation, red = positive correlation). However, there are some that are quite correlated with each other. However, they are quite correlated with each other. For example, HBM, LOC101927069, and PITPNM2.1 are fairly correlated with each other, and H2AFV, KIAA0101, and COL11A1 are quite correlated as well. As shown in Fig. 4, gene correlation of the 24 genes for GSE70947 gene expression includes EZH2.1, COL10A1.1, and LOC100132724, as well as SDPR, LincRNA.chr2.120459730.120511405_R, and KCNA4.1.

4) Deep learning for classification module: The H2O package serves as the foundation for the deep learning technique, and it uses multi-layer neural networks that have been trained using stochastic gradient descent search to forecast the results of diagnoses. To achieve the best classification outcomes for the neural network setup, H2O enables users to conduct hyper parameter grid searches on several deep learning models. Rectifier or Tanh are often the activation functions.

A single hidden layer site (100 or 200 neurons), two discrete layer locations (10, 20 or 50 neurons each), three discrete layers with 30 neurons each, and four discrete layers with 25 neurons each are predefined for assortments. The feed-in dropout ratio options are available in steps of 0.1 from 0 to 0.9.

Typically, there are zero or two total training samples per iteration, where 0 represents one epoch and 2 represents two. The H2O package chooses the automatic value with caution. The maximum number of epochs (iterations) to run the entire dataset is set at 500. Momentum starts out at a value of 0 or 0.5. (Default zero, without hyper-parameter grid search.).

TABLE II. LIST OF SOME OF THE BEST INFORMATIVE GENES FOR GSE70947

gene symbol	SU values
EZH2.1	0.4083618
COL10A1.1	0.4042125
LOC100132724	0.3069758
lincRNA.chr2	0.2469364
MS4A1.1	0.1901143
PTPN1.1	0.1774414
COL1A1.1	0.1766899
TNKS.1	0.1616661
N4BP2L1	0.1605699
BAX.7	0.1580289

TABLE III. LIST OF SOME OF THE BEST INFORMATIVE GENES FOR GSE19804

gene symbol	SU values
COL10A1	0.72990832
PROM2	0.68648556
SH3GL3.1	0.68648556
GOLM1.1	0.68648556
RTKN2	0.64935173
CA4	0.58091266
DPP6.2	0.54387741
FLJ30901....SCUBE1	0.53771981
HS6ST2.2	0.53771981
CNTN6	0.53265203

Full image size

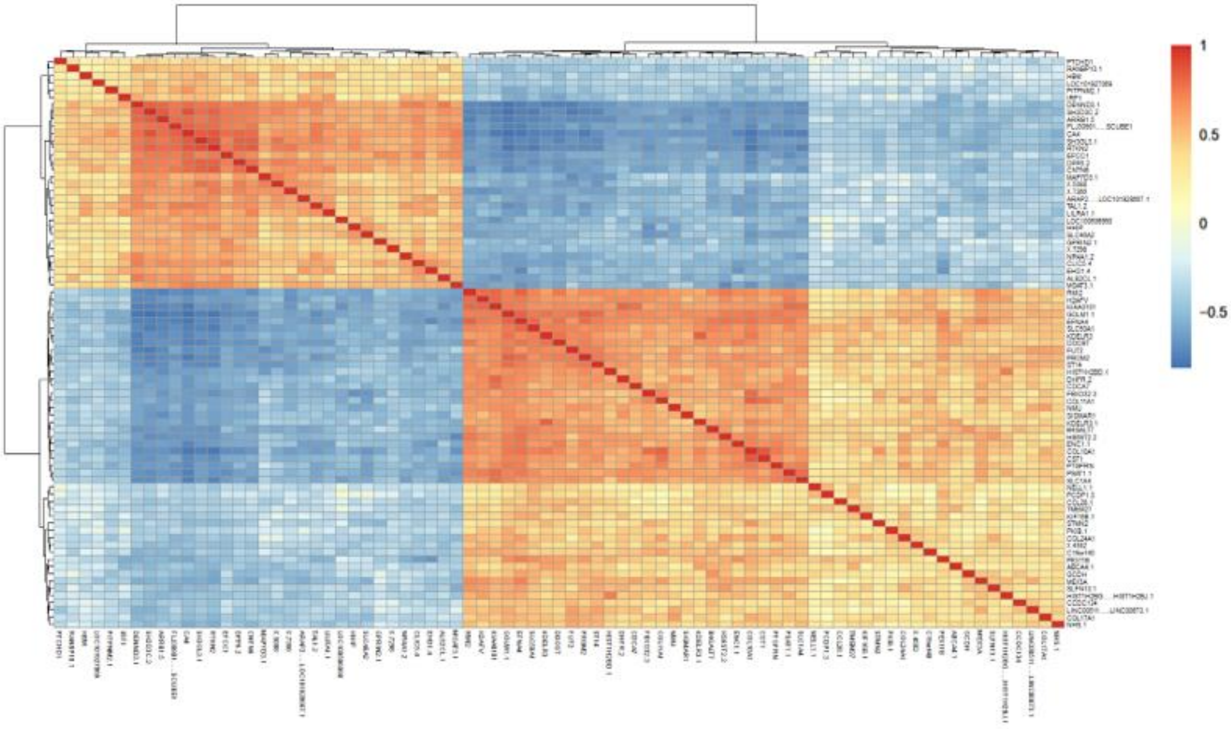

Fig. 3. Heat map for GSE19804 gene expression.

Full image size

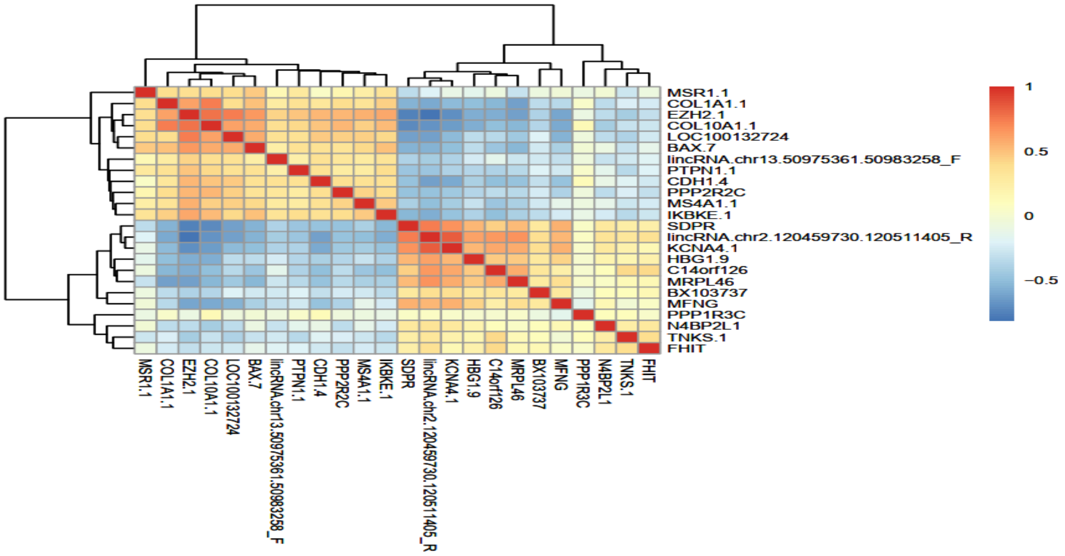

Fig. 4. Heat map for GSE70947 gene expression.

The momentum quickens the iterations for a quicker concourse and dampens the oscillation to achieve the optimal spot. 0.5 Or 0.99 is the adaptive learning rate decay factor (). (Default 0.99, devoid of hyper-parameter grid search) while simultaneously doing quantile regression, the quantile rate (quantile alpha rate in H2O) is set between 0 and 1. Contrary to linear regression, which tests the answer variable's provisional mean, quantile regression tests the provisional quantile. Between 0 and 1 is set as the threshold between quadratic and linear loss (Huber alpha rate in H2O) (default 0.9). In order to make it easier to search on entire combinations of the hyper-parameters, the "random discrete" technique is abandoned.

The most extreme number of models for each run is set at 100 as part of the automatic ML training. If the misclassification values do not increase by 0.01 after five iterations, the training phases come to an end. Score duty cycle, which refers to how frequently validation metrics are computed, is set to 0.025 H2O, which means that no more than 2.5% of the total training time will be spent to create the validation measurements.

Following grid search, the final hyper-parameters for the DL model are listed as follows for the Women's Cancers RNA-Seq dataset: "Rectifier" activation function, four hidden layers with 25 neurons each, insert dropout ratio zero, defaulting training samples each iteration per H2O (value of -2), epoch rate of 430.9, momentum beginning value zero, value of 0.99, quantile regression rate one, and a Huber -value of zero.

Additionally, additional hyper-parameters with an L1 regularization rate of 2.5e-5 and an L2 regularization rate of 2.6e-5 are included. Along with the eleven existing machine learning algorithms that were previously used in H2O for classification, these new DL algorithms are PDA, SVMRadial, GaussprPoly, NB, RF, NN, and glmboost. Based on the data and the size of the sample, to prevent overfitting, N-fold cross-validation with a default N of 10 is feasible. With training data that has been cross-validated 10 times.

To obtain average metrics, we randomly repeated this process ten times. As module number six is illustrated in section VI, bar graphs are used to inform classification metrics such as accuracy, F1 score, area under the curve (AUC) score, precision, sensitivity (SEN), and specificity (SPEC).

V. EXPERIMENT RESULTS AND EVALUATION METRICS WITH TEST DATASET

The proposed architecture was developed using R studio with Bioconductor, with dependencies on the following packages: h2o, dplyr, tidyr, GEOquery, ggplot2, FCBF, pheatmap, devtools, ggbiplot, factoextra, ROCR, limma, psych, caret, foreach, DESeq2, MLSeq, affy, genefilter, hgu133a.db, AnnotationDbi, org, M3C, matrixTests, impute, and gbm The implementation was carried out on a computer server with a core CPU (Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50 GHz, 32 GB RAM) and 64-bit operating system, but it may also work on less powerful machines.

The results indicate that the feedforward neural network algorithm is statistically superior in the metric when compared to other algorithms, with an area under the curve score of 0.982 for RNA-Seq breast adenocarcinoma data and 0.980 for RNA-Seq NSCLC cancer data. Whole evaluation metrics are applied to test datasets, as discussed in Table IV.

So, all evaluation metrics achieved higher rates in breast adenocarcinoma data than in NSCLC cancer data, as plotted in Fig 5. By utilizing a feedforward neural network, this suggested model can assist in the early detection and diagnosis of malignancies in women and, consequently, aid in the formulation of preliminary treatment methods to improve survival. Finally, NSCLC and breast adenocarcinoma cancer may be affected by the top ten possible hub-gene biomarker discoveries.

Biomarkers of lung cancer identified by RNA-non-small-cell sequencing as stated in table V, the top seven choices for differentially expressed genes were found to be shared by all methods. All algorithms found COL10A1 as a common factor, indicating that this gene may be important in NSCLC.

As stated in Table V, the top seven choices for differentially expressed genes were found to be shared by all methods. All algorithms found COL10A1 as a common factor, indicating that this gene may be important in NSCLC. As stated in Table VI, the top four choices for differentially expressed genes were found to be shared by all methods.

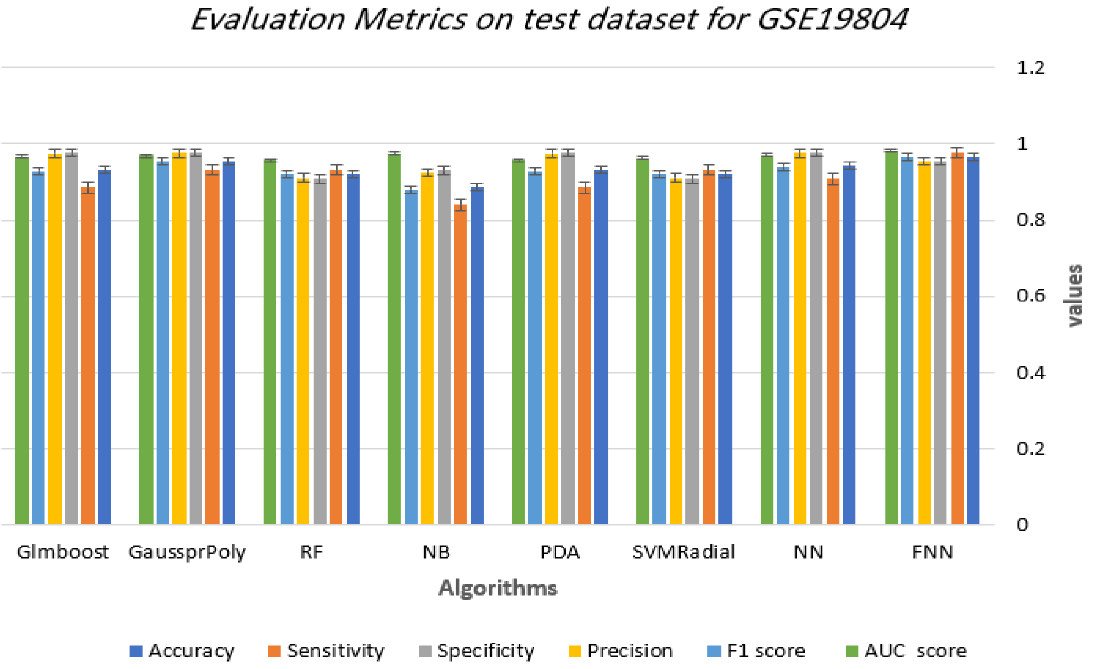

Fig. 5. Evaluation metrics for GSE70947.

TABLE IV. PERFORMANCE EVALUATION OF CLASSIFICATION MODELS ON TEST DATASET

Women's cancers RNA-Seq gene expression Classification using Deep learning with a hybrid Model with Tuning Parameters for Non-Small-Cell Lung Cancer							
Test Datasets	***Algorithm***	***Accuracy***	***Sensitivity***	***Specificity***	***Precision***	***F1 score***	***AUC score***
GSE19804	FNN	0.957	0.971	0.946	0.956	0.965	0.980
	NN	0.9201	0.9316	0.9094	0.9101	0.921	0.9650
	SVMRadial	0.883	0.653	0.927	0.8542	0.891	0.885
	PDA	0.8889	0.7778	1.0000	1.0000	0.875	0.8923
	NB	0.905	0.663	0.920	0.9284	0.907	0.910
	RF	0.892	0.568	0.946	0.8542	0.891	0.877
	GaussprPoly	0.878	0.560	0.939	0.8441	0.882	0.881
	Glmboost	0.906	0.600	0.945	0.9286	0.907	0.911
Women's cancers RNA-Seq gene expression Classification using Deep learning with a hybrid Model with Tuning Parameters for Breast Adenocarcinoma Cancer							
GSE70947	FNN	0.9659	**0.9772**	0.9545	0.9555	0.9662	0.982
	NN	0.9432	0.9091	0.9773	0.9756	0.9412	0.971
	SVMRadial	0.9205	0.9318	0.9091	0.9111	0.9213	0.964
	PDA	0.9318	0.8864	0.9773	0.9750	0.9286	0.957
	NB	0.8864	0.8409	0.9318	0.9250	0.8810	0.975
	RF	0.9205	0.9318	0.9091	0.9111	0.9213	0.957
	GaussprPoly	0.9545	0.9318	0.9773	0.9762	0.9535	0.969
	Glmboost	0.9318	0.8864	0.9773	0.9750	0.9286	0.968

TABLE V. THE BIOMARKERS FOR NON-SMALL-CELL LUNG CANCER RNA-SEQ

SYMBOL	**GENE name**
COL10A1	Collagen Type X Alpha 1 Chain
SH3GL3.1	SH3 Domain Containing GRB2 Like 3, Endophilin A3
GOLM1.1	golgi membrane protein 1
RTKN2	rhotekin 2
EFNA4	ephrin A4
FUT2	fucosyltransferase 2
CLIC5.4	chloride intracellular channel 5

TABLE VI. THE BIOMARKERS FOR BREAST ADENOCARCINOMA RNA-SEQ

SYMBOL	**GENE name**
EZH2.1	enhancer of zeste 2 polycomb repressive complex 2 subunit
COL10A1.1	collagen type X alpha 1 chain
COL1A1.1	collagen type I alpha 1 chain
CDH1.4	cadherin 1

VI. CONCLUSION

Women's cancers are a group of illnesses exhibiting abnormal cell proliferation that have the potential to attack or spread to various body areas. Due to improvements in efficiency and accuracy, RNA-Seq has previously greatly increased the analysis of human genetics and helped to better understand the nature of cancer disorders. In order to classify two different types of cancer, non-small-cell lung cancer and breast adenocarcinoma, this paper introduced an intelligent framework based on a feedforward neural network with an optimization model and applied other integrated learning models suitable for gene expression data for women's cancers. The five modules that made up the suggested strategy were: A core gene identification (such as the Entrez Gene id) is used in the first module, "Feature mapping," which applies an organism-level package (an "org" package) and which contains mappings between this identifier and other types of identifiers (e.g., GenBank or Uniport accession number, etc.). Preprocessing is covered in the second module. The deseq-rlog approach and z-score Normalization and transformation are two techniques used for normalization and transformation. The dimension transformer is the third module. After the RNA-Seq reads have been mapped to a reference genome or transcriptome, the number of reads mapped to the reference genome can be tallied to determine the abundance of the transcripts. For the approaches to be used, it is crucial that the count values be raw sequencing read counts.

The feature selector module, is the fourth module. The fast correlation-based filter (FCBF) was chosen as the method for feature selection in this framework. A deep learning technique, machine predictors with hyper-parameters, and prediction biomarkers with hyper-parameters comprise the final module. Accuracy, sensitivity, specificity, precision, the F1 score, and the area under the curve (AUC) score are among the performance evaluation metrics. Given that the results show that the feedforward neural network approach has an area under the curve score of 0.982 for RNA-Seq breast adenocarcinoma data and 0.980 for RNA-Seq NSCLC cancer data, it is statistically considerably superior to other algorithms in the measure.

References

[1] Saint-Ghislain, Mathilde, Chloé Levenbruck, and Audrey Bellesoeur. "Adverse events of targeted therapies approved for women's cancers." International Journal of Women's Dermatology (2021).

[2] Naz, Faiza, et al. "The role of long non-coding RNAs (lncRNAs) in female oriented cancers." Cancers 13.23 (2021): 6102.

[3] Hamzeh, Osama, and Luis Rueda. "A gene-disease-based machine learning approach to identify prostate cancer biomarkers." Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 2019.

[4] Moffitt, Jeffrey R., Emma Lundberg, and Holger Heyn. "The emerging landscape of spatial profiling technologies." Nature Reviews Genetics (2022): 1-19.

[5] Vimal, Divya, and Khadija Banu. "Developmental Genetics." Genetics Fundamentals Notes. Springer, Singapore, 2022. 955-1027.

[6] Joseph, M., Madhavi Devaraj, and Larry A. Vea. "Cancer classification of gene expression data using machine learning models." 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM). IEEE, 2018.

[7] Xie, Ying, et al. "Early lung cancer diagnostic biomarker discovery by machine learning methods." Translational oncology 14.1 (2021): 100907.

[8] Hong, Mingye, et al. "RNA sequencing: new technologies and applications in cancer research." Journal of hematology & oncology 13.1 (2020): 1-16.

[9] Kaya, S. Irem, et al. "Recent achievements and challenges on nanomaterial based electrochemical biosensors for the detection of colon and lung cancer biomarkers." Sensors and Actuators B: Chemical 351 (2022): 130856.

[10] Mattath, Mohamed Nabeel, et al. "Nucleic Acid Architectonics for pH-Responsive DNA Systems and Devices." ACS omega 7.4 (2022): 3167-3176.

[11] Salmen, Fredrik, et al. "High-throughput total RNA sequencing in single cells using VASA-seq." Nature Biotechnology (2022): 1-14.

[12] Boussios, Stergios, et al. "BRCA mutations in ovarian and prostate cancer: Bench to bedside." Cancers 14.16 (2022): 3888.

[13] Alhenawi, Esra'A., et al. "Feature selection methods on gene expression microarray data for cancer classification: A systematic review." Computers in Biology and Medicine 140 (2022): 105051.

[14] Gunasundari, B., and S. Arun. "Ensemble Classifier with Hybrid Feature Transformation for High Dimensional Data in Healthcare." 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, 2022.

[15] Kong, JungHo, et al. "Network-based machine learning approach to predict immunotherapy response in cancer patients." Nature communications 13.1 (2022): 1-15.

[16] (Zhang, Shihua, et al. "Prediction of lymph-node metastasis in cancers using differentially expressed mRNA and non-coding RNA signatures." Frontiers in Cell and Developmental Biology 9 (2021): 605977.) 17. (Han, Y., Yang, J., Qian, X., Cheng, W. C., Liu, S.

[17] H., Hua, X., ... & Lu, Y. (2019). DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. Nucleic acids research.) 18. Simsek, Ebru, Hasan.

[18] Badem, and Ibrahim Taner Okumus. "Leukemia Sub-Type Classification by Using Machine Learning Techniques on Gene Expression." Proceedings of Sixth International Congress on Information and Communication Technology. Springer, Singapore, 2022.

[19] Das, Ananya, and Subhashis Chatterjee. "Cancer Classification Based on an Integrated Clustering and Classification Model Using Gene Expression Data." *International Conference on Artificial Intelligence and Sustainable Engineering*. Springer, Singapore, 2022.

[20] Yin, Qingyan, et al. "A convolutional neural network model for survival prediction based on prognosis-related cascaded Wx feature selection." Laboratory Investigation 102.10 (2022): 1064-1074. 21.

[21] Houssein, Essam H., et al. "A hybrid barnacle mating optimizer algorithm with support vector machines for gene selection of microarray cancer classification." IEEE Access 9 (2021): 64895-64905. 22.

[22] (Vaiyapuri, Thavavel, et al. "Red Fox Optimizer with Data-Science-Enabled Microarray Gene Expression Classification Model." Applied Sciences 12.9 (2022): 4172.).

[23] Shen, Jiquan, et al. "Deep learning approach for cancer subtype classification using highdimensional gene expression data." BMC bioinformatics 23.1 (2022): 1-17.

[24] Rezaee, Khosro, et al. "Deep learning-based microarray cancer classification and ensemble gene selection approach." IET Systems Biology 16.3-4 (2022): 120-131.

[25] Kolberg, Liis, et al. "gprofiler2--an R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler." F1000Research 9 (2020).

[26] Zhao, Yingdong, et al. "TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models' repository." Journal of translational medicine 19.1 (2021): 1-15.

[27] Bu, Dechao, et al. "KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis." Nucleic acids research 49.W1 (2021): W317-W325.

[28] Almugren, Nada, and Hala Alshamlan. "A survey on hybrid feature selection methods in microarray gene expression data for cancer classification." IEEE access 7 (2019): 78533-78548.

[29] Azadifar, Saeid, et al. "Graph-based relevancy-redundancy gene selection method for cancer diagnosis." Computers in Biology and Medicine 147 (2022): 105766.

[30] Liu, Hongyu, and Bo Lang. "Machine learning and deep learning methods for intrusion detection systems: A survey." applied sciences 9.20 (2019): 4396.

[31] Haldorai, Anandakumar, and Arulmurugan Ramu. "Canonical correlation analysis based hyper basis feedforward neural network classification for urban sustainability." Neural Processing Letters 53.4 (2021): 2385-2401.