A Machine Learning Model for Cancer Disease Diagnosis using Gene Expression Data

Suhaam Adnan Abdul kareem
Department of Postgraduate of Affairs
University of Baghdad
Baghdad, Iraq
Seham.adnan@uobagaghdad.edu.iq
Orcid.org/0000-0002-2584-9946

Zena Fouad Rasheed
Department of Petroleum, College of Engineering
University of Baghdad
Baghdad, Iraq
Zena.fouad@coeng.uobaghdad.edu.iq
Orcid.org/0000-0002-0190-164X

DOI: http://dx.doi.org/10.31642/JoKMC/2018/100227
Received Jun. 22, 2023. Accepted for publication Jul.23, 2023

Abstract— Cancer is one of the top causes of death globally. Recently, microarray gene expression data has been used to aid in cancers effective and early detection. The use of machine learning techniques in biomedicine and bioinformatics to categorize cancer patients into high- or low-risk groups was investigated by numerous research teams. It is necessary that machine learning tools can recognize important features in complex datasets. Here we present a machine learning approach to cancer detection, and to the identification of genes critical for the diagnosis of cancer .We used the Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and Gradient Boosting (GB) that provide results that are more accurate than those of current models. Each model's accuracy, including SVM, KNN, RF, and GB, was (97.41%, 89.3%, 88.1%, and 85.7%), respectively. The SVM has the highest precision among machine learning algorithms. By creating a machine learning-based predictive system for early detection, our findings can help to decrease the prevalence of cancer disease.

Keywords— Cancer, Machine Learning, SVM, Decision tress, Random Forest.

I. INTRODUCTION

The aberrant cell proliferation that characterizes the cancer group of disorders. In a healthy body, cell proliferation is under control, allowing for a predictable pattern of cell growth and degeneration. The genetic makeup of the cells may be harmed by internal and external influences, which causes the cells to continue to proliferate and eventually become tumors [1]. The main causes of cancer are internal factors like improper cell division and damage to deoxyribonucleic acid, while environmental factors like sun exposure, radiation, and chemicals in tobacco smoke play a major role [2][3]. Lung cancer cases have been surpassed by female breast cancer cases and are one of the most often detected forms of cancer. Figure (1) shows the cancer cases and deaths in 2020.

Gene expression profiles are used to diagnose and categorize the cancer condition. Microarray gene expression data processing is one of the difficult study subjects in the fields of Computational Biology, Genomics, Statistics, and Pattern Classification. The main problem with microarray cancer analysis is the short sample size and high curse of dimensionality brought on by redundant and irrelevant genes [3][4] .

Additionally, the majority of medical datasets exhibit noise, varying feature values, and an unbalanced number of classes, all of which contribute to over-fitting and reduced classification accuracy [5][6].

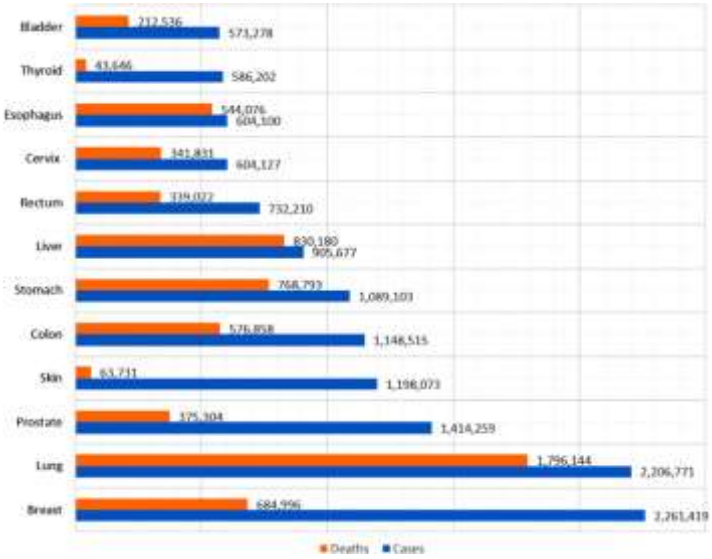


Figure -1 Cancer cases and deaths in 2020 *(3)*

This aids in the early detection of cancer so that specialists in the field can develop a treatment strategy to increase the survival rate of cancer patients [7] [8].

The classification of microarray cancer data involves several key steps, consisting of source data collection, pre-processing, feature selection, classification, and post-classification analysis [9]. When classifying cancer data, feature selection is essential for identifying the best and most pertinent subsets of features and improving classification accuracy and computational

stability [10][11][12].

When attempting to distinguish between several gene expressions profiles, a binary classification or a multiclass classification challenge can appear. Cancer subtypes or malignant and non-cancerous profiles can be used to categorize the profiles. They use random variables (X1, X2, etc.) to represent the expression values of a group of genes (G1, G2, etc.). The gene expression profile of every sample is represented by a tuple Ti. Ti contains a label C and n genes' expression values. The total number of classes in the data is K, and C is also a random variable with a range of K potential values. The profiles are split into training and test sets, with test sets lacking the labels while the training sets do. The difficulty in identifying cancer is in finding a function f that can accept training data with labels as input and predict unknown labels with the highest degree of accuracy.

In this article, we offer a novel feature selection approach and show how well it performs in I differentiating disease samples from healthy samples, (ii) classifying various disease samples, and (iii) identifying disease subtypes. First, we assessed how well our technique classified diseases using a variety of huge cancer gene expression datasets. After that, we compared our findings using the same dataset and other feature/gene set selection techniques. According to the findings, our suggested method can identify crucial genes in diseases generally, and cancer in particular and it performs better than all the well-known gene set selection methods.

The rest of the paper is structured as follows. In Section 2, we go over related scholarly articles. The suggested approach is presented in Section 3. Section 4 presents the experimental design and outcomes analysis. Discussion and comparative analysis are covered in Section 5; the closing thoughts and suggested future research are offered in Section 6.

II. LITERATURE REVIEW

Due to current research initiatives and advancements in biomedical and information technology, a variety of algorithms that are useful for cancer diagnosis using a variety of data-driven diagnostic procedures have been developed [13]. Mabu et al. [14] for the classification of gene expression datasets, an artificial neural network technique and proposed cluster-based feature selection are described, however, classifying pointless photos may reduce its effectiveness. Zeebaree et al. [15] used a convolution neural network to suggest a method of gene selection and categorization using data from cancer microarrays, however, additional samples need to be classified, and the classifier's effectiveness must be increased. Mohapatra et al. [16] the feature weights for the proposed ridge regression (RR) with a single hidden layer feed-forward network were generated at random. Binary microarray datasets for breast, prostate, colon tumor, and leukemia were used to validate the approach. It has been noted that the dataset for breast cancer does not follow the typical train/test routine properly.

The Sharbaf et al. [17] a hybrid approach for gene selection and classification of microarray datasets has been developed, integrating cellular learning automata with ant colony optimization, when the huge nodule annotations were insufficient, the efficiency was, however, constrained. Joyseeree. They employ three classifiers, support vector machine, (KNN), and Naive Bayes for validation, and they have studied the effects of various rank-based feature selection techniques. Kumar et al.[18] developed feature selection and classification techniques using MapReduce in combination with the KNN classifier, however, it must classify other illnesses or tumors. Nguyen et al. [19] they created an aggregate gene selection for microarray data categorization and tested their model using the four widely used datasets DLBCL, Leukemia, Prostate, and Colon, however, it uses a small number of data points to demonstrate its effectiveness. Five well-known classifiers, including linear discriminant analysis, KNN, probabilistic neural network, SVM, and Multilayer Perceptron (MLP), were used to validate the suggested method, but the stability beyond these five classifiers could not be proven.

III. Data Collection

The value of the suggested methodology was illustrated using a collection of three publicly available data sets that included cancer sample microarray data and had no missing data. The National Center for Biotechnology Information's Gene Expression Omnibus (GEO) was used to download data sets with the identifiers GSE65194 and GSE 20711, which were obtained using Affymetrix Human Genome U133 Plus 2.0 arrays, and GSE25055, which was obtained using Human Genome HG U133A Affymetrix arrays (NCBI).

IV. Methodology

A four supervised learning classifiers used in this study are as follows: Using the Python "Sklearn" package, different kernels of the Support Vector Machine, (RF), (KNN), and (GB) models were developed. The data set contains information on the cancer cells, enabling this information to be used as inputs and matching outputs, which is why the supervised learning models were chosen. When given classification tasks with labeled data, supervised models do especially well. Before implementing any algorithms, data cleansing was done. The data set was cleaned up of duplicate or incomplete information, and further checks were made to make sure the data was accurate. Users should check, for instance, that all data for each attribute was within the acceptable ranges (1-10). For training, testing, and validation purposes, the data was divided into parts. To compare the effectiveness of the models, analytics for each model were acquired, including accuracy,F1, precision, and recall scores. Data utilized for testing and training were split 80:20. Figure (2) shows how this procedure is visualized.

The categorization of cancer using gene expression data is shown in Figure (3). Data on gene expression is gathered, and the gene's encoded information causes the protein molecule to be produced in the gene expression. The data is moved to the preprocessing stage, where it is improved in quality and formatted in an understandable manner. Principal component analysis is used to minimize the dimension of the vast amount of preprocessed data. The data are sent to the feature selection process after preprocessing. The selected features are gathered by the SVM classifier, which then categorizes the labeled data. Finally, the display presents the overall impression of the classified data.

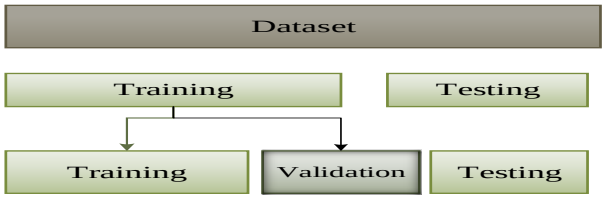


Figure -2 Visualization of data partition segments

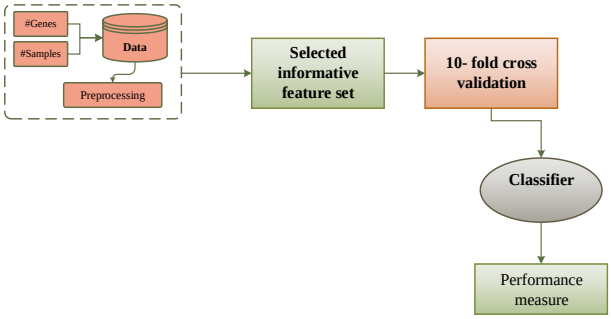


Figure -3 Block diagram of the cancer classification using gene expression data

A. Data Pre-processing

The accuracy and effectiveness of the model are constrained by the real data, which contains sounds, unsupported formats, and lost data. To ensure the quality of the data, gene expression data should be treated beforehand. Cleaning, integrating, reducing, transferring, and discretizing data are steps in the preprocessing of data. To improve quality, the cleaning procedure entails identifying and removing errors. The quality of the data is decreased by incorrect spellings or inaccurate data. Gathering and merging varied data from numerous databases is the process of integration. When the amount of data is very great, data reduction occurs, and occasionally it analyzes the data that is most suitable from a variety of data types. The transformation process involves modification and accumulation, depending on the data requirements. Data discretization entails separating the theoretical properties from the statistical ones.

B. Feature Selection

We just used a portion of the features in our algorithm to carry out the classification. In this case, we chose k characteristics to preserve the most patient data possible. We minimized the classification's over-fitting by using gene-selection, which increased accuracy. We saw the dataset as a collection of patient pairings, much like the feature-selection problem [18] does. If there is a trait that is up-regulated or down-regulated in both individuals, then two patients are said to be comparable.

C. Cancer Classification using Support Vector Machine

When a machine learning model is trained on an already-existing dataset that has been labeled with the expected output, this process is known as supervised learning or model training [19]. As an illustration, categorizing individuals who have previously received a diagnosis of a particular condition using medical data. This is done to aid the model's ability to understand the connections between the dataset's variables and attributes. The test data would be used without labels to predict the desired result after the model had been trained on the training set of data. Model training is the process through which a model learns from previous input-output examples and applies that knowledge to new inputs in order to predict future outcomes [20]. The typical flow of supervised machine learning (model training) is shown in Figure (4) below [20].

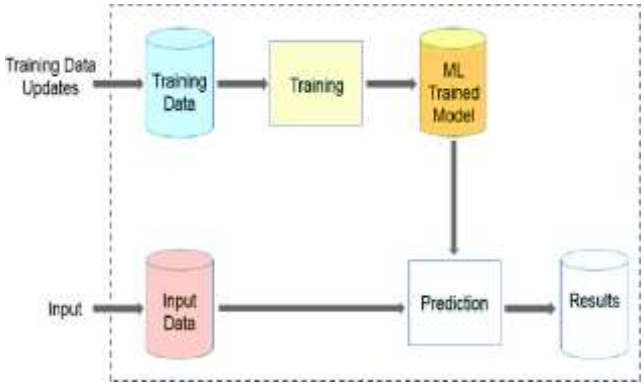


Figure -4 Supervised learning workflow [17]

The SVM the process of creating algorithms for a priori determined categories is known as supervised classification, often known as prediction or discrimination. To assess the correctness of algorithms, they are often constructed on a training dataset and then tested on a separate test dataset. Support vector machines are a collection of associated supervised learning techniques used for regression and classification tasks [20].

SVM maps data to a high-dimensional feature space, which enables the categorization of data even when the data cannot be separated linearly. A separator between the various categories is identified and then the data is transformed in a manner that the separator can be drawn as a hyper-plane. The SVM selects the extreme vectors or points that assist in the creation of the hyper-plane [21]. These extreme cases are known as support vectors and therefore the algorithm is called the support vector machine. In Figure (5), two separate categories are grouped based on a decision boundary hyper-plane.

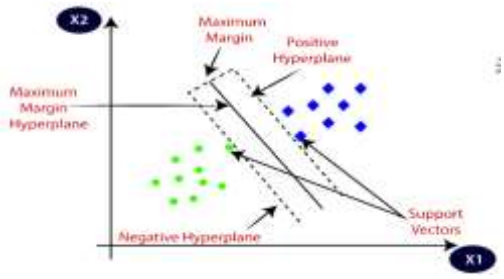


Figure -5 The hyper-plane source [22]

D. Performance Model

Different measures are used to assess the performance of classifier algorithms. True Positives are labels that the algorithm correctly predicts will be positive. True Negatives occur when the model correctly predicts the sample's negative

class that is; when both the actual class labels and the predicted class labels are negative. False Positives are labels that the algorithm predicts will be positive even when they are actually negative. False Positives are labels that the algorithm predicts will be positive even when they are actually negative. For binary classification, n is equal to 2, and the confusion matrix is an n*n table that shows the connection between the actual and anticipated labels for each of the n classes. The algorithm achieves training convergence as the training loss and validation loss level off [21].

1-Accuracy: The percentage sample in the source data set that were correctly assigned to all of the samples in the data set is known as classification accuracy. A classifier output known as (TP) is one in which both the prediction and the actual class are positive. The (TN) is an output from a classifier in which the predicted class is also a true negative. When the classifier predicts a positive outcome when the true class is a negative one, this is known as a false positive (FP) or false negative (FN) classification error [21]. In Eq. (1), accuracy is evident.

$$\text{Acc} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

2-Precision: A precision of a classifier's predictions is defined as the proportion of true positives to all positives[21]. Eq. (2) defines precision.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

3-Recall: The recall of a classification algorithm is the proportion of true positives to the sum of true positives and false negatives. It shows a percentage of real positives that a classifier properly identified[21]. In Eq.(3) recall is defined .

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

4-F1-Measure: Precision and recall are harmonically represented by the F1-score[21]. Score F1 is described in Eq. (4).

$$F-measure = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

V. Results and Discussion

This section presents and discusses the experiment's results. When a proposed feature selection strategy was applied to the datasets and the classifiers were trained using the genes that had been previously chosen from each dataset, we saw very positive performance metrics across the board. When compared to other classifiers, only KNN performance was worse. The SVM is compared with several other machine learning models including SVM with default parameters' values. These models are (RF), (KNN), and (GB). Table (1) below presents the default parameters used for executing the other models. It is important to mention that the experiments are executed.

Table 1 Parameters setting for classification models

Model	Parameter	Value
RF	No. of Trees	50
	Limit Depth	5
KNN	K	5
GB	No. of Trees	100
	Learning rate	0.3
	Lambda	10
SVM	C	1000

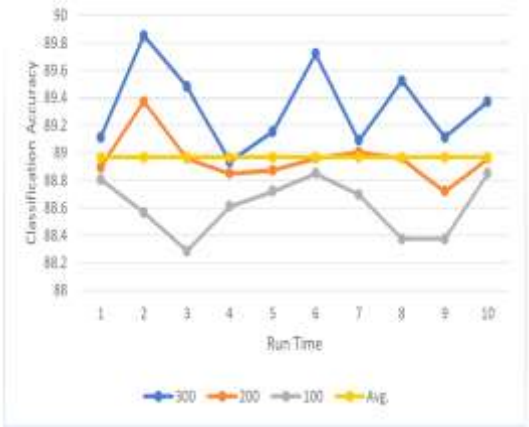
It can be seen from (Table 2) that the proposed SVM algorithm has a very good performance for tuning the hyper-parameters of SVM and keeps the searching process stable. The stability of the algorithm is proven by checking the standard deviation value of the algorithm, which is always low and close to zero, which means SVM for 10 times has almost reach near results with no big difference. In general, the mean accuracy shows that SVM reached very good accuracy for all 10 run times as compared to standard SVM with default parameters' values.

Figure (6) (C-D) illustrates a full and clear picture on all 10 runtimes for the algorithms .Table (2): Classification Performance of the Proposed Method.

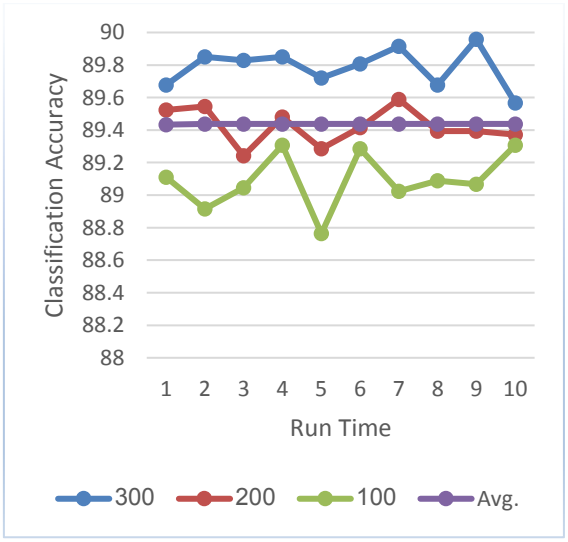
Table 2 Performance of SVM on two datasets

Dataset	SS	MaxItr	Best Acc %	F1%	Precision %	Recall%
GSE20711	10	100	84.09	77.56	80.01	80.92
		200	86.17	79.29	81.00	82.69
		300	88.10	80.00	82.10	84.93
	20	100	90.54	81.05	84.30	89.57
		200	93.00	82.00	85.20	91.13
		300	87.00	80.61	81.19	83.97
	30	100	89.10	81.02	83.60	85.09
		200	93.09	83.51	86.70	90.73
		300	96.41	84.00	88.00	93.09

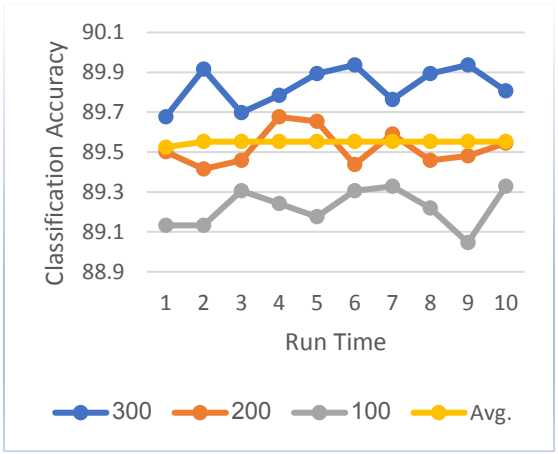
GSE65194	10	100	83.01	80.00	81.00	82.10
		200	85.3	72.10	81.25	83.02
		300	89.12	73.10	82.50	84.25
	20	100	90.21	76.11	83.37	89.02
		200	92.21	79.10	85.89	90.00
		300	86.07	82.12	81.21	84.20
	30	100	90.10	74.00	83.02	86.36
		200	91.08	80.00	85.10	89.18
		300	94.18	83.00	87.00	90.21



C) Results obtained when SS = 10, MaxItr = {100, 200, 300}



D) Results obtained when SS = 20, MaxItr = {100, 200, 300}



E) Results obtained when SS = 30, MaxItr = {100, 200, 300}

Figure -6 Results obtained using different configurations for 10 different runtimes

The performance of SVM is compared against several well-known machine learning models. Table 3 below presents a full comparison between the models. It is clear that our proposed model have attained better classification accuracy as compared to other models while maintaining other metrics such as recall and precision, which means SVM to find better values for (C &γ), and decreases the chances of over-fitting.

Table 3 The adopted measurement criteria for performance evaluation

Dataset	Model	F1	Precision	Recall	Accuracy
GSE 20711	RF	80.3%	80.5%	81.1%	88.1%
	KNN	74.1%	76.6%	77.4%	85.7%
	GB	80.2%	82.6%	83.3%	89.3%
	SVM	82.3%	85.2%	85.7%	97.41%
GSE65194	RF	80.5%	81.3%	84.1%	86.01%
	KNN	77.1%	79.6%	80.1%	84.2%
	GB	80.5%	80.6%	82.3%	88.3%
	SVM	88.3%	90.01%	91.5%	94.18%

Figures (7) bellows illustrate the convergence of SVM when the number of iterations is set to 50.

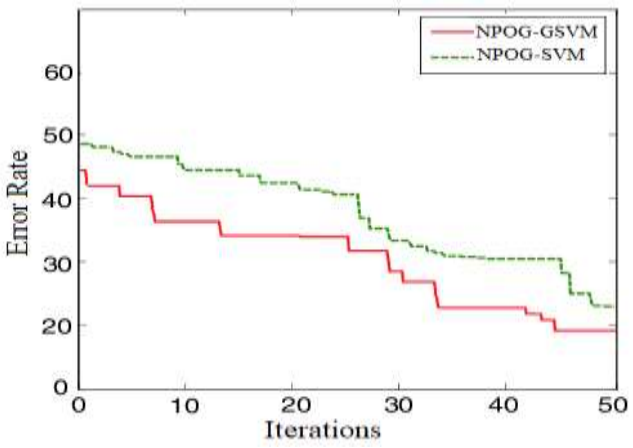


Figure -7 Convergence analysis of the proposed gene selection algorithms

VI. Conclusion

To use machine learning techniques on high-dimensional cancer gene expression datasets, sample size issues must be fixed, choosing the relevant gene subset, class imbalance, and unstable results. In our study, we have addressed these difficulties. We scaled the data to produce a mean of 0 and a variation of 1. This stops a single trait with a significant variance across samples from being the dominant one in the findings. The dataset's class distribution was balanced using synthetic minority oversampling. Then, derived from the initial gene-rich data sets, a small number of important genes were chosen. Four classifiers SVM, Random Forest, K-NN, and GB, were trained using chosen genes. With practically all of the data sets and all classifiers, we saw very good performance numbers. The suggested method has only been tested on microarray data in the current work; however, via evaluating it using data from different sources as well as data from next-generation sequencing, this constraint can be overcome in the future.

References

[1] K. Hall, V. Chang, and P. Mitchell, "Machine Learning Techniques for Breast Cancer Detection," no. January, pp. 116–122, 2022.

[2] S. T. Ahmed and S. M. Kadhem, "Early Alzheimer's Disease Detection Using Different Techniques Based on Microarray Data: A Review," *International journal of online and biomedical engineering*, vol. 18, no. 4, pp. 106–126, 2022.

[3] S. T. Ahmed and S. M. Kadhem, "Using Machine Learning via Deep Learning Algorithms to Diagnose the Lung Disease Based on Chest Imaging: A Survey," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 16, pp. 95–112, 2021.

[4] S. T. Ahmed and S. M. Kadhem, "Alzheimer's disease prediction using three machine learning methods," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 3, pp. 1689–1697, 2022.

[5] H. S. M. Alsultani, S. T. Ahmed, B. J. Khadhim, and Q. K. Kadhim, "The use of spatial relationships and object identification in image understanding," *International Journal of Civil Engineering and Technology*, vol. 9, no. 5, pp. 487–496, 2018.

[6] S. T. Ahmed, Q. K. Kadhim, H. S. Mahdi, and W. S. A. Almahdy, "Applying the MCMSI for Online Educational Systems Using the Two-Factor Authentication," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 13, pp. 162–171, 2021.

[7] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91–107, 2017.

[8] A. Bir-Jmel, S. M. Douiri, and S. Elbernoussi, "Gene selection via a new hybrid ant colony optimization algorithm for cancer classification in high-dimensional data," *Computational and mathematical methods in medicine*, vol. 2019, 2019.

[9] H. S. Basavegowda and G. Dagnew, "Deep learning approach for microarray cancer data classification," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 22–33, 2020.

[10] M. Mandal, P. K. Singh, M. F. Ijaz, J. Shafi, and R. Sarkar, "A tri-stage wrapper-filter feature selection framework for disease classification," *Sensors*, vol. 21, no. 16, p. 5571, 2021.

[11] Saroj, J. Vashishtha, P. Goyal, and J. Ahuja, "A Novel Fitness Computation Framework for Nature Inspired Classification Algorithms," *Procedia Computer Science*, vol. 132, pp. 208–217, 2018.

[12] Y. Saeys, "Inza, I. n.; and Larranaga," *A review of feature selection techniques in bioinformatics. Bioinformatics*, vol. 23, no. 19, pp. 2507–2517.

[13] K. Mukesh, "nitish, KR, Amitav, S., & Santanu, KR (2015). Feature Selection and Classification of Microarray Data Using Map Reduce Based ANOVA and K-Nearest Neighbor, 11th International Multi-Conference on Information Processing," *Procedia Computer Science*, vol. 54, pp. 301–310.

[14] N. Q. K. Le, D. T. Do, T.-T.-D. Nguyen, N. T. K. Nguyen, T. N. K. Hung, and N. T. T. Trang, "Identification of gene expression signatures for psoriasis classification using machine learning techniques," *Medicine in Omics*, vol. 1, no. May 2020, p. 100001, 2021.

[15] Y. Tian and Z. Qi, "Review on: Twin Support Vector Machines," *Annals of Data Science*, vol. 1, no. 2, pp. 253–277, 2014.

[16] A. N. Parveen, H. H. Inbarani, and E. N. S. Kumar, "Performance analysis of unsupervised feature selection methods," in *2012 International Conference on Computing, Communication and Applications*, 2012, pp. 1–7.

[17] R. Liu, C. A. Mancuso, A. Yannakopoulos, K. A. Johnson, and A. Krishnan, "Supervised learning is an accurate method for network-based gene classification," *Bioinformatics*, vol. 36, no. 11, pp. 3457–3465, 2020.

[18] A. Masood *et al.*, "Computer-Assisted Decision Support System in Pulmonary Cancer detection and

stage classification on CT images," *Journal of Biomedical Informatics*, vol. 79, no. January, pp. 117–128, 2018.

[19] H. Akkar and S. Q. Haddad, "Diagnosis of Lung Cancer Disease Based on Back-Propagation Artificial Neural Network Algorithm," *Engineering and Technology Journal*, vol. 38, no. 3B, pp. 184–196, 2020.

[20] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," *Computers in Biology and Medicine*, vol. 140, p. 105051, 2022.

[21] C. M. Rosett and A. Hagerty, "Introducing Machine Learning," in *Introducing HR Analytics with Machine Learning*, Springer, 2021, pp. 107–127.

[22] Z. Mao, W. Cai, and X. Shao, "Selecting significant genes by randomization test for cancer classification using gene expression data," *Journal of biomedical informatics*, vol. 46, no. 4, pp. 594–601, 2013.