Indonesian Journal of Electrical Engineering and Computer Science
Vol. 24, No. 3, December 2021, pp. 1610~1617
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v24.i3.pp1610-1617

Parallel extreme gradient boosting classifier for lung cancer detection

Rana Dhiaa Abdu-Aljabar, Osama A. Awad
Faculty of Information Engineering, Al-Nahrain University, Baghdad, Iraq

Article Info

Article history:

Received Jun 29, 2021
Revised Oct 27, 2021
Accepted Nov 1, 2021

Keywords:

Bioinformatics
Gene expression
Lung cancer disease
Machine learning
XGBoost

ABSTRACT

Most lung cancers do not cause symptoms until the disease is in its later stage. That led the lung cancer having a high fatality rate compared to other cancer types. Many scientists try to use artificial intelligence algorithms to produce accurate lung cancer detection. This paper used extreme gradient boosting (XGBoost) models as a base model for its effectiveness. It enhanced lung cancer detection performance by suggesting three stages model; feature stage, XGBooste parallel stage and selection stage. This study used two types of gene expression datasets; RNA-sequence and microarray profiles. The results presented the effectiveness of the proposed model, especially in dealing with imbalanced datasets, by having 100% each of sensitivity, specificity, precision, F1_score, area under curve (AUC), and accuracy metrics when it applied on all of the datasets used in this study.

Corresponding Author:

Rana Dhiaa Abdu-Aljabar
Department of Systems Engineering, Faculty of Information Engineering, Al-Nahrain University
Baghdad, Iraq
Email: ranadhiaa1@gmail.com

1. INTRODUCTION

Lung cancer is common cancer that causes a higher fatality rate between cancer types. The five-year survival rate is about 56% for patients that cancer is still in the lung. While 5% for the cases, its cancer spread out of the lung. Only 16% of lung cancer cases are detected early [1]. Recognition and prediction the lung cancer in the earliest stage can increase the survival rate of the patients. Lung cancer has no symptoms in the early stages [2], [3], so it needs more than traditional detection to detect it. Cancer can be defined as a disease of altered gene expression. The development of gene expression technologies has become the standard technology for studying the cells [4]-[6]. The development of this technology made many researchers apply many studies on improving lung cancer prediction by analyzing the changes in gene expression. Some researchers study gene expression-based prognostic signatures for lung cancer [3]. Others try to use gene expression technology such as microarray and RNA-sequences to develop lung cancer detection methods. Many studies used artificial intelligence to detect lung cancer for their power tools. They used different methods, like Al-Anni *et al.* [7]-[12], they proposed different optimization models to enhance the non-small cell lung cancer (NSCLC) detection accuracy using microarray gene expression datasets. Also, Azzawi *et al.* [13]-[16] have multiple studies on multiclass improvement using the GEP algorithm in the lung cancer classification stage to determine the specific therapy and reduce the fatality rate. Hu *et al.* [17] proposed detecting and recognizing different life stages of bladder cells using two cascaded convolutional neural networks (CNNs). To detect cancer cells and their stages. While Saric *et al.* [18] proposed a fully automatic method for detecting lung cancer in lung tissue. They used two convolutional neural network CNN architectures (VGG and ResNet) for training, and their performance is compared. The results obtained show that the CNN-based approach can help pathologists diagnose lung cancer. Also, Li *et al.* [19] proposed a fusion algorithm that combines handcrafted features into the features learned at the output layer of a 3D deep convolutional neural network (CNN). Patra [20] analyzed various machine learning classifier techniques to classify lung cancer into benign and malignant. Lai *et al.* [21] trained clinical and gene expression data with improved deep neural network (DNN). It used patients based on microarray data to predict the 5-year survival status of NSCLC. The study Priya and Jawhar [22] proposed an automatic approach to classifying the lung image into a normal case or cancer case by preprocessing the computed tomography (CT) lung image to remove noise. Then combines the histogram analysis with morphological and extracts the lung regions by thresholding operations, while Ogunleye and Wang [23] used a clinical database to classify the patient if he has chronic kidney disease or not using XGBoost. Desuky *et al.* [24] suggested a new method for classification to deal with imbalanced medical datasets. It used the crossover to increase the minority class and then used the boosting for classification. This method enhances the classification accuracy results. Rustam *et al.* [25] used feature extraction from CT images as data to classify lung cancer. Fuzzy C-Means and fuzzy kernel C-Means were used to classify the lung nodule from the patient into benign or malignant. The score showed fuzzy kernel C-Means had higher accuracy than fuzzy C-Means accuracy. Pandian *et al.* [26] developed an algorithm to classify lung cancer medical images as normal and infected. The features are extracted from CT images of normal lung and cancer affected lungs were taken into the study. The artificial neural network is used in classification. Hakim *et al.* [27] compared two popular feature selection models to enhance the support vector machine (SVM) cancer classification. They showed that the ReliefF outperformed compared with CFS as microarray data feature_selection approach. Kareem *et al.* [28] developed the CT scanning data set using imaging/computer vision algorithms for diary of healthy and tumorous chest scans; This comprises three preprocessing steps: i) improvement of images, ii) segmentation of images, and iii) strategies for feature extraction. In the last stage, a support vector machine (SVM) is utilized to categorize slide instances as one of 3 types (normal, benign, or malignant) by using classification technology. The best accuracy, 89,8876 percent, was obtained when applying this technique to the new dataset. Selwal and Raoof [29] developed a MATLAB-based CNN for automated detection of cancerous cervix cells where the templates segmented the nucleus of the cells. The simulation results show that the proposed CNN algorithm can automatically detect the cervix cancer cells with more than 88% accuracy. Raju *et al.* [30] used high resolution computer tomography (HRCT) images with multi-classification to classify 17 interstitial lung diseases with convolutional neural network (CNN) architecture called SmallerVGGNet. It obtained 95% averaged accuracy. Ali *et al.* [31] deep neural networks were used, that is, the enhancer Deep Belief Network (DBN), which is constructed from two Restricted Boltzmann Machines (RBM). The enhancer DBN was trained by back propagation neural network (BPNN). It found that LASSO with LR gives the best accuracy in their study dataset. Abdullah *et al.* [32] developed a MATLAB-based CNN for automated detection of cancerous cervix cells where the templates segmented the nucleus of the cells. The proposed CNN algorithm detected the cervix cancer cells automatically with more than 88% accuracy.

A previous study [33] compares multiple current machine learning and found that the XGBoost is the most accurate system in balance and imbalance datasets. This study tried to improve the XGBoost by using a feature selection to use only the genes responsible for lung cancer disease in the learning stage and applied a parallel XGBoost (PXGB) with different hyperparameters to increase the system variety and decrease the overfitting and underfitting. The PXGB showed more accurate prediction values for detecting cancer and normal lung state, especially for imbalanced datasets.

2. XGBOOST ALGORITHM

XGBoost is a decision-tree-based ensemble machine learning algorithm developed by Tianqi Chen and Carlos Guestrin. They employ machine learning algorithms under the gradient boosting framework. They introduced their work at the SIGKDD conference in 2016 [34].

First, let us clarify the concept of boosting. It is an ensemble method that seeks to create a robust classifier (model) based on "weak" classifiers. In this context, weak and robust refer to how correlated the learners are to the actual target variable. By adding models on top of each other iteratively, the errors of the previous model are corrected by the next predictor, sequentially in the training stage, as it appeared in figure1 until the training data is accurately predicted or reproduced by the model. Finally, as seen in Figure 1, it provides a parallel tree boosting for the testing stage that quickly and accurately solves many data science problems. It offers a range of hyperparameters that give fine-grained control over the model training procedure. It worked very well with the imbalance database. It had many features suitable for large databases like; parallelization, distributed computing, out-of-core computing (for managing the large dataset with the memory), cache optimization (to the best use of hardware) [35].

3. LUNG CANCER DATASETS

The datasets used in this study are microarray and RNA-sequence datasets. The data gathered through microarrays represents the gene expression profiles, which show simultaneous changes in the expression of many genes in response to a particular condition or treatment. They represent the molecular level states of the cell [6]. RNA-sequence datasets used a sequencing technique (next-generation sequencing) to disclose the presence and quantity of RNA in a biological sample at a given moment, analyzing the continuously changing cellular transcriptome [36]. This study applied the proposed model on two microarray datasets and one RNA-sequence dataset as shown in Table 1. All datasets were downloaded from the gene expression omnibus site (GEO).

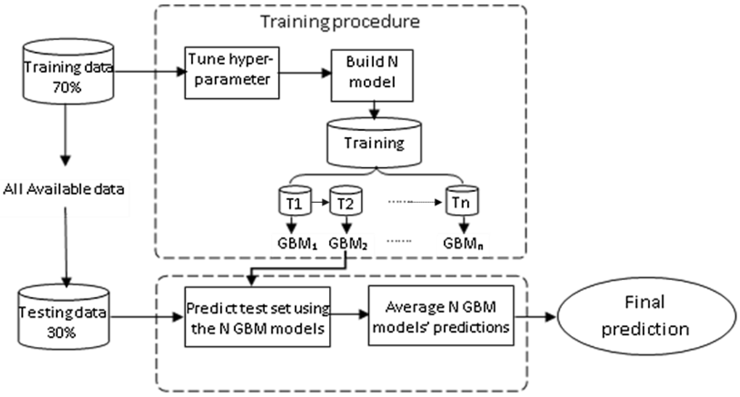

Figure 1. XGBoost algorithm [33]

3.1. Dataset information

Each dataset used has a different way of extracting the gene expression, the number of features and the number of cases. The first is (GSE30219) dataset representing the gene expression by microarray technology. It has 14 normal lung samples and 293 lung cancer samples [37]. The second (GSE74706) dataset is also represented by microarray technology. It expresses data of early-stage NSCLC. It has 18 lung cancer samples and 18 normal lung samples. The last dataset (GSE81089) [38] has 218 cases expressed by RNA-sequencing, which is called next-generation sequencing [39]; RNA-Seq allows researchers to detect gene fusions variants, both known and novel features and other features without the limitation of prior knowledge [40]. It has 199 lung cancer samples with lung cancer type NSCLC and 19 healthy lung samples.

Table 1. Dataset's information

Datasets	Type	patients	Features	The Class	Sample distribution	
					Cancer case	Normal case
GSE30219	Microarray	307	54675	Cancer/Normal	293	14
GSE74706	Microarray	36	34182	Cancer/Normal	18	18
GSE81089	New Generation Sequencing (NGS)	218	63129	Cancer /Normal	199	19

3.2. Data preprocessing

Data preprocessing in machine learning is an essential step in enhancing data quality to raise meaningful perceptiveness. It refers to cleaning and organizing the raw data to make it suitable for building and training machine learning models. In biological data, it is crucial to clean the data to improve the quality of the data for searching and analyzing. To do that, it runs a process to detect and remove corrupt or inaccurate records from the database. Each record with missing data must be deleted because it is regarded as irrelevant and cause inappropriate learning results. The XGBoost classification deals with the numeric representation in the decision class. In contrast, the classes in the lung cancer datasets are in nominal representation, like normal/cancer. Therefore, it must change them to numeric representation (0/1).

4. THE PARALLEL_XGBOOST (PXGB)

There is no way to teach one machine learning to fit all kinds of information. In our case, the XGBoost succeeded in learning on some datasets with high accuracy but lower in others. That is because of its firm reliance on its hyperparameter setting. This study developed an XGBoosts structure to accommodate different types of datasets by connecting multiple numbers of XGBoosts on parallel with various values of hyperparameters. Then it takes the maximum probability for its prediction, as shown in Figure 2. All the XGboosts are working in parallel not to cause a delay in learning time. As seen in Figure 2, the proposed methodology has three stages:

i) Feature selection stage: The benefit of using XGBoost in feature selection is that after the boosted trees are constructed, they will retrieve the importance scores for each feature. The importance score refers to how useful or valuable each feature was in constructing the model boosted decision trees. The more feature is used, the higher its importance score. This importance is computed for every feature in the dataset, allowing the ranking and comparison between them.

   The importance of every decision tree is estimated by calculating the number of observations responsible for each feature split and increasing the measurement of performance. In every decision tree in the model, the attribute imports are then averaged [23]. In this paper, the importance score threshold setting was (10^{-6}). Each attribute less than this threshold will be neglected. The features of GSE30219, GSE74706, and GSE81089 datasets were (54675), (34182) and (63129), respectively, but after the feature selection stage, it becomes (20), (1) and (8) features.

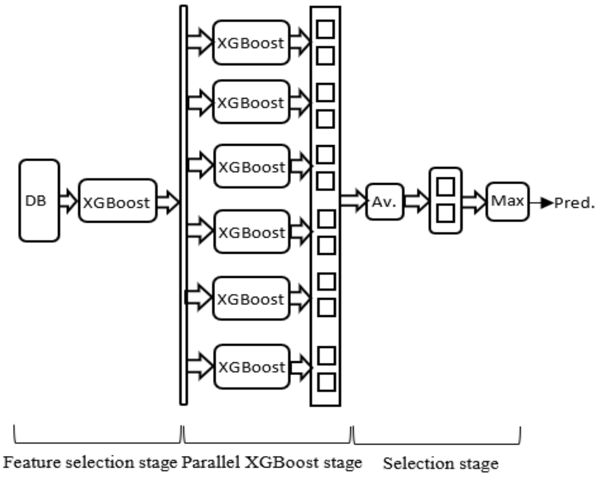

Figure 2. The proposal learning model (PXGB)

ii) Parallel XGBoost stage: After the feature selection stage, the data will be subset to 70% for training and 30% for testing, then entered into each XGBoost simultaneously. In our case, it needs to use different types of bio_dataset. This dataset is usually noisy, so it needs the model to tune its hyperparameters each time to avoid overfitting or underfitting to handle a wide range of datasets. For that reason, It used multi XGBoost models connected in parallel. Each XGBoost has its hyperparameters setting different from each other. This study will take six sets of XGBoost hyperparameters from the most common range that consider the XGBoost model often works well in them. The hyperparameters ranges are; the subsample [0.5 -1], the Max_depth [2-7], the learning rate [0.05-0.3], the n_estimators (no. of trees) [5-50], and the last is the min_child weight from [1-6]. Their arrangement depends on the most values that do not cause overfitting but may sometimes cause an underfitting (level one), to the more values that may cause overfitting but less likely causing underfitting (level six), see Table 2. The end of this stage will have a probability prediction for both classes in each level.

iii) Selection stage: At this stage, it will take the maximum probability value of all XGBoost levels. The result is that the class with maximum probability is chosen as the final class prediction.

Table 2. The setting of each XGBoost hyperparameters in the PXGB

XGBoost sequence	XGBoost hyperparameters				
in the parallel stage	subsamble	Max_depth	Learning rate	n_estimators	min_child_weight
First level	0.5	2	0.3	5	6
Second level	0.6	3	0.25	10	5
Third level	0.7	4	0.2	20	4
Fourth level	0.8	5	0.15	30	3
Fifth level	0.9	6	0.1	40	2
Sixth level	1	7	0.05	50	1

5. THE RESULTS AND DISCUSSION

To state the findings of the research, it arranged the results in a logical sequence. At first, it compared the result of the PXGB model with the original XGBoost [35] to represent its improvement. It then compared its result with representative machine learnings to give the whole performance state of the proposal.

5.1. XGBoost hyperparameters setting

The PXGB sets seven common hyperparameters for each XGBoost classification model, as illustrated in Table 2. These hyperparameters are chosen depending on their arrangement from most hyperparameters that may cause the overfitting situation to the most parameters that may cause the underfitting situation. The setting of the original XGBoost and SVM [41], deep forest (gcforest) [42], k-nearest neighbors algorithm (KNN), and naive Bayes, which are the machine learnings that used in this study, have a particular setting illustrated in Table 3.

Table 3. Parameters setting of representative models

XGBoost		SVM		gcForest		KNN		Naive Bayes	
Parameter	value	Parameters	value	Parameter	value	Parameter	value	Parameter	value
max_depth	6	kernel	RBF	max_depth	6	n_neighbor	2	var_smoothing	1e-9
n_estimators (Trees)	2	gamma	1	no. of trees in each forest stages= 500	500	weights	uniform	sample_weight	None
Learning rate	0.3	tolerance	0.001	Wind. size	500	algorithm	auto		
min_child_weight	1	C	1	Step	100	leaf_size	1		
Subsample	0.7			Min_samples_split	0.7				

5.2. Comparison of different classifiers

Different results were obtained after applying the PXGB model and other machine learning models to the lung cancer datasets. Tables 4 illustrate each model's sensitivity, specificity, precision, F1_score, area under curve (AUC), accuracy, and learning time metrics. Furthermore, Figures 3, 4, and 5 show the receiver operating characteristic (ROC) drawings and the AUC values of each machine learning model.

Table 4. Comparison results of lung cancer detection for all dataset

GSE81089 dataset						
Classifier Name	Sensitivity	Specificity	Precision	F1_score	AUC	Accuracy	Time (min.)
PXGBS	1.0	1.0	1.0	1.0	1.0	1.0	00:03
XGBoost	1.0	1.0	1.0	1.0	1.0	1.0	00:04
SVM	0.2	0.83	0.5	0.29	0.52	0.55	00:01
gcForest	1.0	0	0.45	0.63	0.50	0.45	00:36
KNN	0.8	1.0	1.0	0.89	0.90	0.91	00:01
Naive Bayes	0.6	0.67	0.6	0.6	0.63	0.64	00:01
Classifier Name	Sensitivity	Specificity	Precision	F1_score	AUC	Accuracy	Time (min.)
PXGBS	1.0	1.0	1.0	1.0	1.0	1.0	00:13
XGBoost	1.0	0.95	1.0	0.99	0.99	0.98	00:24
SVM	1.0	0.5	0.95	0.98	0.75	0.95	00:05
gcForest	0.98	0.83	0.98	0.98	0.91	0.97	03:37
KNN	0.95	0.5	0.95	0.95	0.72	0.91	00:29
Naive Bayes	1.0	0.17	0.92	0.96	0.58	0.92	00:02
GSE74706 dataset							
Classifier Name	Sensitivity	Specificity	Precision	F1_score	AUC	Accuracy	Time (min.)
PXGBS	1.0	1.0	1.0	1.0	1.0	1.0	00:13
XGBoost	0.99	1.0	0.99	1.0	0.99	0.99	00:17
SVM	1.0	0	0.96	0.98	0.5	0.96	00:07
gcForest	0.98	0.75	0.98	0.98	0.87	0.97	03:26
KNN	0.98	1.0	1.0	0.99	0.99	0.99	00:12
Naive Bayes	0.99	1.0	1.0	0.99	0.99	0.99	00:02

5.3. Analyzing metrics

From Table 4, it is seen that all PXGB metrics have excellent values when applied to all datasets. It succeeded in detecting all cases (cancer and normal cases) in all datasets. In contrast, XGBoost successfully predicts all cases only in GSE81089 dataset because it has only one set of hyperparameters, while XGBoost has a range of hyperparameters that let it build multiple XGBoost structures in the training stage. PXGB gives the flexibility to deal with different datasets and allows all the XGBoost structures to contribute to the class detection in the test stage and then choose the best prediction by selecting the class with the maximum prediction value.

The PXGB improved the performance of the XGBoost. It has become more powerful and reliable for a variant type of dataset without changing its hyperparameters. Although the naive Bayes has the shortest learning time in most datasets, the PXGB has an accepted learning time ranging from 3 to 13 seconds. It is even shorter than the original XGBoost ranging from 4 to 23 seconds because of the selection feature process, and the multiple XGBoost are worked in parallel, decreasing the system overhead.

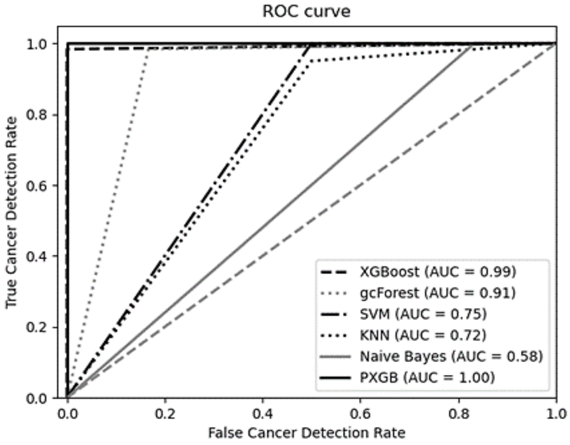

Figure 3. The ROC curves and AUC values for all comparative models on GSE81089 dataset

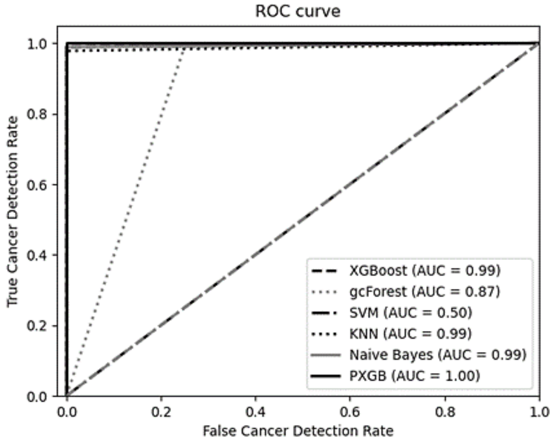

Figure 4. The ROC curves and AUC values for all comparative models on GSE30219 dataset

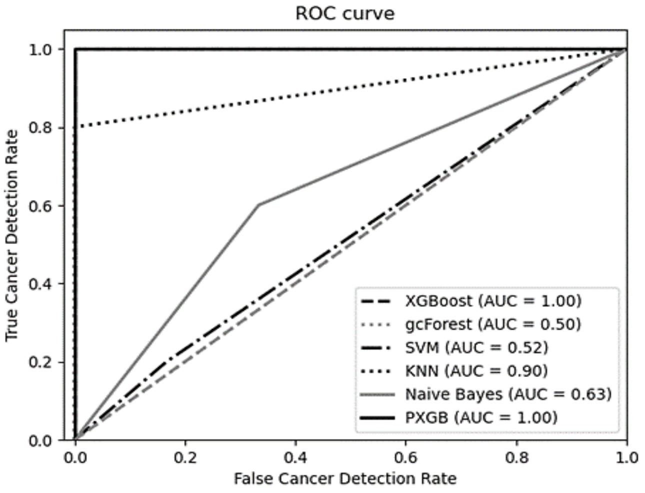

Figure 5. The ROC curves and AUC values for all comparative models on GSE74706 dataset

6. CONCLUSION

This study proposed a lung cancer detection system with multi-stages to reach optimal results. It uses the XGBoost as a feature selection to choose only active genes that have an essential role in lung cancer disease and suggested a flexible machine learning by using multiple XGBoost classifications to run in parallel. Each XGBoost in parallel stage has different sets of hyperparameters, ranging from the most values that may lead to overfitting to the parameters' values that might cause the underfitting. That led to obtaining various tree buildings, which gives the PXGB flexibility and reliability; when applied to different datasets. Moreover, using feature selection improved the detection accuracy; it also sped up the learning time. The results showed that the PXGB model, the proposed model, improved lung cancer detection performance. This improvement is better than the original XGBoost and other comparative machine learning, especially for imbalanced datasets and within an acceptable time.

REFERENCES

[1] World Health Organization, Cancer, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer.

[2] R. Park, J. W. Shaw, A. Korn, and J. McAuliffe, "The value of immunotherapy for survivors of stage IV non-small cell lung cancer: patient perspectives on quality of life," *J. Cancer Surviv,* vol.14, no. 3, pp. 363-376, 2020, doi: 10.1007/s11764-020-00853-3.

[3] Y. Wang *et al*., "A novel 4-gene signature for overall survival prediction in lung adenocarcinoma patients with lymph node metastasis Cancer," *Cell Int.*, vol 19, no. 100, 2019, doi: 10.1186/s12935-019-0822-1.

[4] E. F. Nuwaysir, M. Bittner, J. Trent, J. C. Barrett, and C. A. Afshari, "Microarrays and toxicology: the advent of toxicogenomics," *Molecular Carcinogenesis*, vol. 24, no. 3, pp. 153-159, 1999, doi: 10.1002/(sici)1098-2744(199903)24:3%3C153::aid-mc1%3E3.0.co;2-p.

[5] Y. Yang, E. A. G. Blomme, and J. F. Waring, "Toxicogenomics in drug discover: From preclinical studies to clinical trials," *Chem. Biol. Interact*, vol. 150, no. 1, pp. 71-85, 2004, doi: 10.1016/j.cbi.2004.09.013.

[6] H. A. Rueda-Zárate, I. Imaz-Rosshandler, R. A. Cárdenas-Ovando, J. E. Castillo-Fernández, J. Noguez-Monroy, and C. Rangel-Escareño, "A computational toxicogenomics approach identifies a list of highly hepatotoxic compounds from a large microarray database," *Plos One*, vol. 12, no. 4, 2017, doi: 10.1371/journal.pone.0176284.

[7] R. Al-Anni, J. Hou, R. D. Abdu-Aljabar, and Y. Xiang, "Prediction of NSCLC recurrence from microarray data with GEP," *IET systems biology*, vol. 11, no. 3, pp. 77-78. 2017, doi: 10.1049/iet-syb.2016.0033.

[8] R. Al-Anni, J. Hou, H. Azzawi and Y. Xiang, "Cancer adjuvant chemotherapy prediction model for non-small cell lung cancer", *IET systems biology*, vol. 13, no. 3, pp. 129-135, 2019, doi: 10.1049/iet-syb.2018.5060.

[9] R. Al-Anni, J. Hou, H. Azzawi, and Y. Xiang, "Risk classification for NSCLC survival using microarray and clinical data," *International Journal of Advances in Electronics and Computer Science*, vol. 6, no. 3, 2019.

[10] R. Al-Anni, J. Hou, H. Azzawi and Y. Xiang, "A novel gene selection algorithm for cancer classification using microarray datasets," *BMC Med. Genomics*, vol. 12, no. 10, 2018, doi: 10.1186/s12920-018-0447-6.

[11] R. Al-Anni, J. Hou, H. Azzawi and Y. Xiang, "Deep gene selection method to select genes from microarray datasets for cancer classification," *BMC-informatics*, vol. 20, no. 608, 2018, doi: 10.1186/s12859-019-3161-2.

[12] R. Al-Anni, J. Hou, H. Azzawi and Y. Xiang, "New Gene Selection Method Using Gene Expression Programing Approach on Microarray Data Sets," *Int. Conf. on Comp. and Info. Science 4th*, vol. 791, pp. 17-31, 2019, doi: 10.1007/978-3-319-98693-7_2.

[13] H. Azzawi, J. Hou, Y. Xiang and R. Alanni, "Lung cancer prediction from microarray data by gene expression programming," *IET Syst. Biol*., vol. 10, no. 5, pp. 168-178, 2016, doi: 10.1049/iet-syb.2015.0082.

[14] H. Azzawi, J. Hou, Y. Xiang, R. Alanni, R. Abdu-aljabar and A. Azzawi, "Multiclass lung cancer diagnosis by gene expression programming and microarray datasets," *13th Int. Conf. on Advanced Data Mining and Applications*, vol. 38, 2017, pp 541-553, doi: 10.1007/978-3-319-69179-4_38.

[15] H. Azzawi, J. Hou, R. Alnnni, and Y. Xiang, "SBC: A New Strategy for Multiclass Lung Cancer Classification Based on Tumour Structural Information and Microarray Data," *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, 2018, pp. 68-73, doi: 10.1109/ICIS.2018.8466448.

[16] H. Azzawi, J. Hou, R. Alanni, and Y. Xian, "A hybrid neural network approach for lung cancer classification with gene expression dataset and prior biological knowledge," *Int. Conf. on Machine Learning for Networking*, pp. 279-293, 2019, doi: 10.1007/978-3-030-19945-6_20.

[17] H. Hu, Q. Guan, S. Chen, Z. Ji and Y. Lin, "Detection and Recognition for Life State of Cell Cancer Using Two-Stage Cascade CNNs," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 3, pp. 887-898, 1 May-June 2020, doi: 10.1109/TCBB.2017.2780842.

[18] M. Šarić, M. Russo, M. Stella, and M. Sikora, "CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images," *2019 4th International Conference on Smart and Sustainable Technologies (SpliTech)*, 2019, pp. 1-4, doi: 10.23919/SpliTech.2019.8783041.

[19] S. Li *et al.*, "Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features Physics in Medicine & Biology," *Physics in Medicine and Biology*, vol. 64 no. 17, 2019, doi: 10.1088/1361-6560/ab326a.

[20] R. Patra, "Prediction of Lung Cancer Using Machine Learning Classifier," *Int. Conf. on Computing Science, Communication and Security Computing Science, Communication and Security*, pp. 132-142, 2020, doi: 10.1007/978-981-15-6648-6_11.

[21] Y.-H. Lai, W.-N. Chen, T.-C. Hsu, C. Lin, Y. Tsao, and S. Wu, "Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning," *Scientific Reports*, vol. 10, no. 4679, 2020, doi: 10.1038/s41598-020-61588-w.

[22] M. M. A. Priya and S. J. Jawhar, "Advanced lung cancer classification approach adopting modified graph clustering and whale optimisation-based feature selection technique accompanied by a hybrid ensemble classifier," *IET Image Processing*, vol. 14, no. 10, pp. 2204-2215, 2020, doi: 10.1049/iet-ipr.2019.0178.

[23] A. Ogunleye and Q.-G. Wang, "XGBoost Model for Chronic Kidney Disease Diagnosis," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 6, pp. 2131-2140, 1 Nov.-Dec. 2020, doi: 10.1109/TCBB.2019.2911071.

[24] A. S. Desuky, A. H. Omar, and N. M. Mostafa, "Boosting with crossover for improving imbalanced medical datasets classification," *Bulletin of Electrical Engineering and Informatics,* vol. 10, no. 5, pp. 2733-2741, 2021, doi: 10.11591/eei.v10i5.3121.

[25] Z. Rustam, A. Purwanto, S. Hartini, and G. S. Saragih, "Lung cancer classification using fuzzy c-means and fuzzy kernel C-Means based on CT scan image," *Indonesian Journal of Applied Informatics*, vol. 10, no. 2, pp. 291-297, 2021, doi: 10.11591/ijai.v10.i2.pp291-297.

[26] R. Pandian, D. N. S. R. Kumar, and R. R. Kumar, "Development of algorithm for identification of maligant growth in cancer using artificial neural network", *International Journal of Electrical and Computer Engineering,* vol. 10, no. 6, pp. 5709-5713, 2020, doi: 10.11591/ijece.v10i6.

[27] M. A. N. Hakim, Adiwijaya, and W. Astuti, "Comparative analysis of ReliefF-SVM and CFS-SVM for microarray data classification," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, pp. 3393-3402, 2021, doi: 10.11591/ijece.v11i4.pp3393-3402.

[28] H. F. Kareem, M. S. AL-Huseiny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, "Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS),* vol. 21 no. 3, pp. 1731-1738, 2021, doi: 10.11591/ijeecs.v21.i3.pp1731-1738.

[29] A. Selwal and I. Raoof, "A Multi-layer perceptron based intelligent thyroid disease prediction system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 1, pp. 524-532. 2020, doi: 10.11591/ijeecs.v17.i1.pp524-532.

[30] N. Raju, H. B. Anita, and P. Augustine, "Identification of interstitial lung diseases using deep learning", *International Journal of Electrical and Computer Engineering*, vol. 10, no. 6, pp. 6283-6291, 2020, doi: 10.11591/ijece.v10i6.pp6283-6291.

[31] N. M. Ali, N. A. A. Aziz, and R. Besar, "Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 2, pp. 712-719, 2020, doi: 10.11591/ijeecs.v20.i2.pp712-719.

[32] A. A. Abdullah, A. F. D. Giong, and N. A. H. Zahri, "Cervical cancer detection method using an improved cellular neural network (CNN) algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no.1, pp. 210-218, 2019, doi: 10.11591/ijeecs.v14.i1.pp210-218.

[33] R. D. Abdu-aljabar and O. A. Awad, "A Comparative analysis study of lung cancer detection and relapse prediction using XGBoost classifier," *IOP Conf. Ser. Mater. Sci. Eng*., vol. 1076, no. 1, 2021, doi: 10.1088/1757-899X/1076/1/012048.

[34] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *In Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[35] S. S. Dhaliwal, A-A. Nahid, and R. Abbas, "Effective Intrusion Detection System Using XGBoost," *Information*, vol. 9, no. 7, 2018, doi:10.3390/info9070149.

[36] Y. Chu and D. R. Corey, "RNA sequencing: platform selection, experimental design, and data interpretation," *Nucleic Acid Therapeutics*, vol. 22, no. 4 pp: 271-274, 2012, doi: 10.1089/nat.2012.0367.

[37] S. Rousseaux *et al.,* "Ectopic activation of germline and placental genes identifies aggressive metastasis-prone Lung cancers," *Science Translational Medicine*, vol. 22, no. 5, 2013, doi: 10.1126/scitranslmed.3005723.

[38] A. Mezheyeuski *et.al*, "Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify Lung cancer patients," *J Pathol*, vol. 244, no. 4, pp. 421-431, 2018, doi: 10.1002/path.5026.

[39] D. C. Bell *et al*., "DNA base identification by electron microscopy," *Microsc Microanal*, vol. 18, no. 5, pp. 1049-1053, 2012, doi: 10.1017/S1431927612012615.

[40] F. Ozsolak and P. M. Milos, "RNA sequencing: Advances, challenges and opportunities," *Nat Rev Genet.*, vol. 12, pp. 87-98, 2011, doi: 10.1038/nrg2934.

[41] L. Wang, *"Support Vector Machines: Theory and Applications*," Berlin, Germany: Springer, 2005, doi: 10.1007/b95439.

[42] Z-H. Zhou and J. Feng, "Deep Forest: towards an alternative to deep neural networks," *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 3553-3559, doi: 10.24963/ijcai.2017/497.