Parallel Extreme Gradient Boosting Classifier for Lung Cancer Detection

Abstract

Most lung cancers do not cause symptoms until the disease in its later stage. That led the lung cancer has a high fatality rate comparing 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 its 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, AUC, *and accuracy for all of the datasets used.*

Keywords*: Machine learning, Extreme Gradient Boosting, Gene expression, Lung cancer disease, Bioinformatics.*

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 define as a disease of altered gene expression. The gene expression technologies development has become the standard technology for study the cells [4-6]. The development of this technology made many researchers applied 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 and had a good result, like Russul A. et al. [7-12]. They proposed different studies of new optimization models to improve NSCLC detection using microarray gene expression datasets. Also, Hasseeb A. et al.[13-16] have an improvement to multiclass using GEP algorithm in lung cancer classification stage to determine the specific therapy and reduce the fatality rate. Haigen 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 Matko Š., et al. [18] they 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, Shulong 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 R. [20] analyzed various machine learning classifiers techniques to classify lung cancer into benign and malignant.

Lai, Y., 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 of Michael M. A. P. [22] proposed an automatic approach to classifying the lung image into a normal case or cancer case by pre-processing the CT lung image to remove noise. Then combines the histogram analysis with morphological and extracts the lung regions by thresholding operations, while Adeola Ogunleye's study [23] used a clinical database to

classify the patient if he has chronic kidney disease or not using XGBoost. Azian A., et. al. [24] suggested an enhanced cellular neural network (CNN) as a solution for detecting malignant cells in real-time using Pap smear images after image processing. Rozlini Mohamed, et. al [25] used the Bat Algorithm and K-Means techniques for classification performance improvement, which they applied on 14 datasets. Results show that BkMDFS outperforms most performance measures, and they show that Bat Algorithm has the potential to be one of the discretization techniques and feature selection techniques. In a previous study [26], we compare 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 applied a parallel XGBoost (PXGB) with different hyperparameters to increase the system variety and decrease the overfitting. 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 was developed by Tianqi Chen and Carlos Guestrin. They implement machine learning algorithms under the Gradient Boosting framework (see figure 1). They introduced their work at SIGKDD conference in 2016 [27]. XGBoost provides a parallel tree boosting 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.

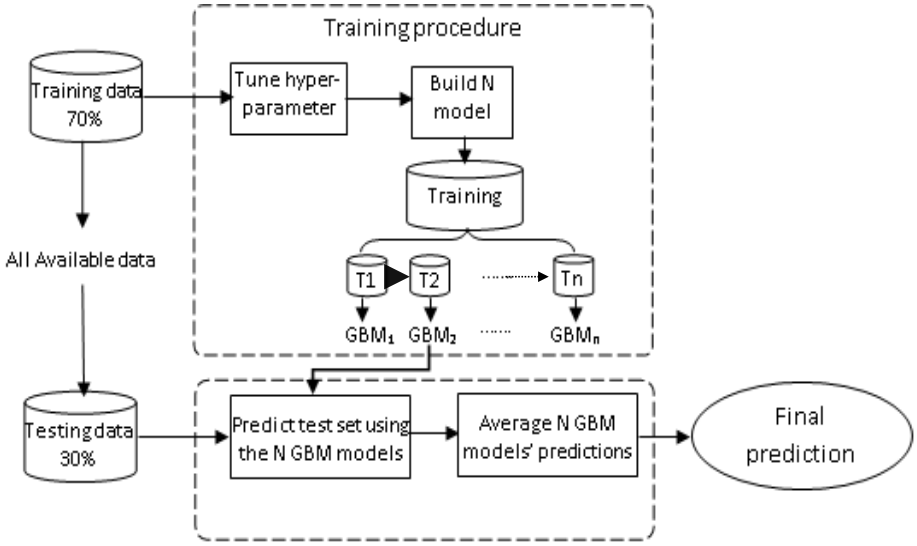

Figure 1. XGBoost algorithm

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.[28]

This study applied the proposed model on two microarray datasets and one RNA-sequence dataset (see Table 1). All datasets were downloaded from the National Center for Biotechnology Information site (NCBI).

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 [29]. The second (GSE74706) dataset is also represented by microarray technology. It is expressing data of early-stage NSCLC. It has 18 lung cancer samples and 18 normal lung samples. The last dataset (GSE81089) [30] has 218 cases expressed by RNA-sequencing, which is called next-generation sequencing [31]; RNA-Seq allows researchers to detect gene fusions variants, both known and novel features, and other features without the limitation of prior knowledge [32]. It has 199 lung cancer samples with NSCLC type and 19 normal 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 pre-processing

Data pre-processing 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 have lower accuracy in others. This because of its firm reliance on its hyperparameter setting. This study made development on XGBoosts structure to accommodate different types of datasets by connecting multiple numbers of XGBoosts on parallel with a variety value of hyperparameters. Then it takes the maximum probability for its prediction, as shown in Fig.2. All the XGboosts are working in parallel not to cause a delay in learning time. As seen in Fig.2, the proposed methodology has three stages:

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 calculated for each feature in the dataset, allowing features to be ranked and compared.

The importance is calculated for each decision tree by counting each feature split point and improving the performance measure, weighted by the number of observations the node is responsible for. The attribute importances are then average across all decision trees within the model [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.

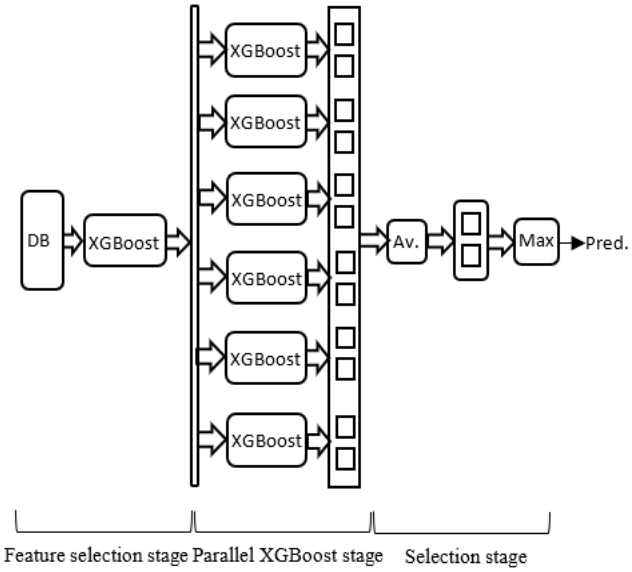

Figure 2. The proposal learning model (PXGB)

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 kind of 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 not caused 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. At the end of this stage, it will have a probability prediction for both classes in each level.

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 in the parallel stage	XGBoost hyperparameters				
	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

The PXGB compare its result with original XGBoost, 2016 [27], support vector machine (SVM), 2005 [33] deep forest (gcforest), 2017 [34], KNN (k-nearest neighbors algorithm) and Naive Bayes.

5.1 XGBoost hyperparameters setting

The PXGB sets the hyperparameters of all XGBoosts as shown in Table 2, and each of the original XGBoost, SVM, gcforest, KNN, and Naive Bayes have a particular setting, as shown in Table 3.

Table 3. Parameters setting of representative models

XGBoost		SVM		gcForest		KNN		Naive Bayes	
Parameter	value	Parameter	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	500	weights	uniform	sample_weight	None
Learning rate	0.3	tolerance	0.001	Wind. size	500	algorithm	auto		
min_child_weigh	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, AUC, accuracy, and learning time metrics. Furthermore, in figures 3, 4, and 5, they showed the 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
GSE30219 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	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 applying 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. Despite the Nave Byse has

the shortest learning time in most datasets, the PXGB has an accepted learning time ranged 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.

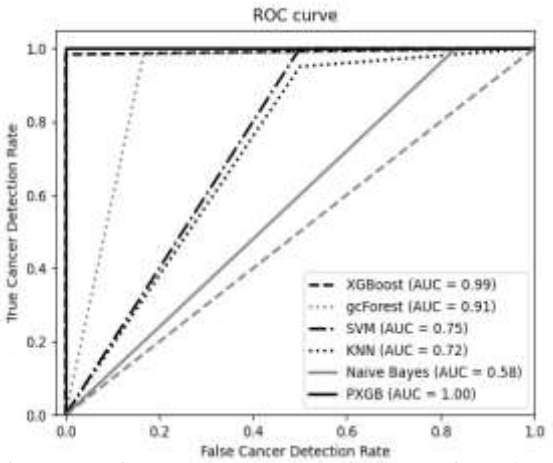

Figure 3. The ROC curves and AUC values for all comparative models on GSE81089 dataset.

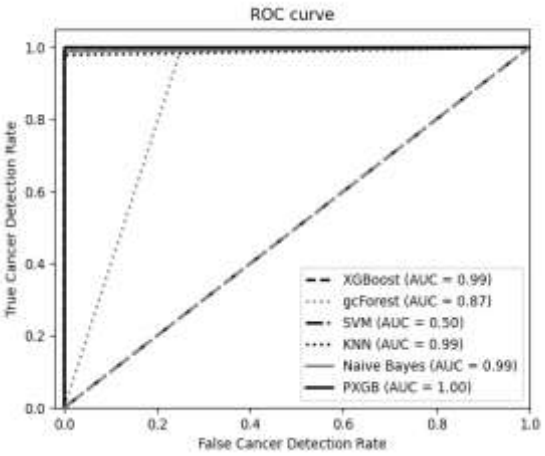

Figure 4. The ROC curves and AUC values for all comparative models on GSE30219 dataset.

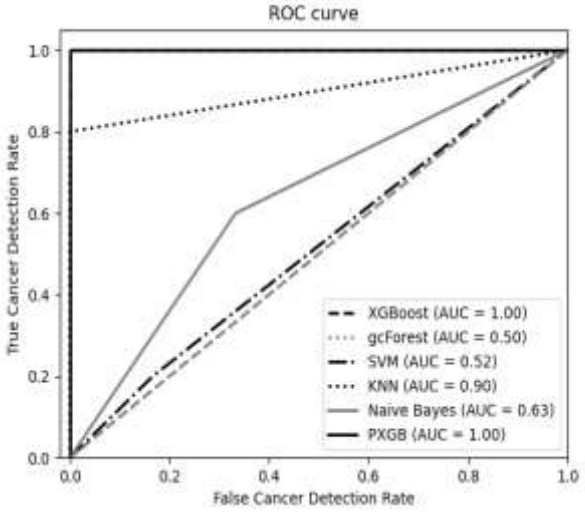

Figure 5. The ROC curves and AUC values for all comparative models on GSE74706 dataset

6. Conclusions

This study suggested flexible machine learning for lung cancer detection using multiple XGBoost classifications connected in parallel. Each XGBoost has different hyperparameter ranges from the most values that from learning relations might be led to overfitting to the values that might cause the underfitting, to obtain various tree buildings. This variety gives the PXGB flexibility and reliability when applied to different datasets. Also, using the XGBoost algorithm as a feature selection to the PXGB model improved its accuracy and sped up the learning time.

The results showed that the PXGB 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 Fact Sheet, 2021.

[2] R. Park, J. W. Shaw, A. Korn, et. el. The value of immunotherapy for survivors of stage IV non-small cell lung cancer: patient perspectives on quality of life, *J Cancer Surviv*, 2020; 14(3): 363–376.

[3] Y. Wang, Q. Zhang, Z. Gao, S. Xin, Y. Zhao, K. Zhang, R Shi and X. Bao, 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.

[4] E. F. Nuwaysir, M Bittner, J. Trent, C A Afshari. Microarrays and toxicology: the advent of toxicogenomics, *Molecular Carcinogenesis*, 1999; 24(3): 153-159.

[5] Yi YangEric, A.G. BlommeEric, A.G. BlommeJeffrey, F WaringJeffrey and F Waring, Toxicogenomics in drug discover: From preclinical studies to clinical trials. *Chem. Biol. Interact*, 2004;150(1): 71-85.

[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*, 2017; 12(4).

[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*, 2017; 11(3): 77-78.

[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*. 2018; 13(3).

[9] R. Al-Anni, J. Hou, H. Azzawi, and Y. Xiang. *Risk classification for NSCLC survival using microarray and clinical data*, Proc. of 207th The IIER Int. Conf. 12th-13th December Paris France. 2018.

[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*, 2018; 12(10).

[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, 2018; 20(608).

[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 Sep. 2018, Springer, Cham., 2019; 791: 7-31.

[13] H. Azzawi, J. Hou, Y. Xiang and R. Alanni, Lung cancer prediction from microarray data by gene expression programming", *IET Syst. Biol*., 2016; 10(5): 168-178.

[14] H. Azzawi, J. Hou, Y. Xiang and R. Alanni, R. D, 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 14 Oct, 2017 , Singapore Springer, Cham. chapter, 2017; 38: pp 541-553..

[15] H. Azzawi, J. Hou, Y. Xiang and R. Alanni. SBC: A new strategy for multiclass Lung cancer classification based on tumour structural information and microarray data, 17th IEEE/ACIS Int. Conf. on Computer and Information Science 6-8 June 2018, Singapore IEEE, 2018; 68-73.

[16] H. Azzawi, J. Hou, Y. Xiang and R. Alanni, *A hybrid neural network approach for lung cancer classification with gene expression dataset and prior biological knowledge*, Int. Conf. on Machine Learning for Networking May 2019 Paris France Springer, Cham Lecture Notes in Computer Science, 2019; 11407: 279-293.

[17] Haigen Hu, Qiu Guan, Shengyong Chen, et. al, Detection And Recognition For Life State Of Cell Cancer Using Two-Stage Cascade Cnns, IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2020; 17(3).

[18] Matko Šarić, Mladen Russo, Maja Stella, et. al., *CNN-based Method for Lung Cancer Detection in Whole Slide Histopathology Images*, 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech) IEEE, 2019.

[19] S. Li, P. Xu, B. Li, L. Chen, Z. Zhou, H. Hao, Y. Duan, M Folkert, J Ma1, S. Huang, S. Jiang and J. Wang Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features Physics in Medicine & Biology. 2019; 64(17).

[20] R. Patra, Prediction of Lung Cancer Using Machine Learning Classifier, Int. Conf. on Computing Science, Communication and Security Computing Science, Communication and Security Springer Singapore, 2020; 1235: 132-42.

[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, Sci. Rep. Nature research, 2020; 10(4679).

[22] M. M. A. Priya 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*. 2020; 14(10): 2204 – 2215.

[23] A. Ogunleye, Q-G. Wang, XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Trans*

Comput Biol Bioinform. 2020; 17(6): 2131-2140.
[24] Azian Azamimi Abdullah1 Aafion Fonetta Dickson Giong, Nik Adilah Hanin Zahri, Cervical cancer detection method using an improved cellular neural network (CNN) algorithm. Indonesian Journal of Electrical Engineering and Computer Science. April 2019; 14(1): 210-218.
[25] Saouabi Mohamed, Abdellah Ezzati, A data mining process using classification techniques for employability prediction, Indonesian Journal of Electrical Engineering and Computer Science. 2019; 14(2): 1025-1029.
[26] R.D Abdu_aljabar,. O.A Awad, *A Comparative analysis study of lung cancer detection and relapse prediction using XGBoost classifier*, IOP Conf. Ser.: Mater. Sci. Eng. 2021;1076(1).
[27] 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; 785–794.
[28] Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*, 2012; 22(4): 271–274.
[29] Sophie Rousseaux, Alexandra Debernardi, Baptiste Jacquiau, A-L. Vitte, A. Vesin, H. Nagy-Mignotte, D. Moro-Sibilot, P-Y Brichon, S. Lantuejoul, P. Hainaut, J. Laffaire, A. d. Reyniès, D. G Beer, J-F. Timsit, Ch. Brambilla, E. Brambilla, S. Khochbin, Ectopic activation of germline and placental genes identifies aggressive metastasis-prone Lung cancers, *Science Translational Medicine*, 2013; 5(186).
[30] A. Mezheyeuski, C. Holst Bergsland, M. Backman, D. Djureinovic, T. Sjöblom, J. Bruun and P. Micke, Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify Lung cancer patients, *J Pathol*, 2018; 244(4): 421-431.
[31] D.C Bell, W K. Thomas, K.M. Murtagh, Ch.A Dionne, A.C Graham, J.E Anderson and W.R Glover, DNA base identification by electron microscopy, *Microsc Microanal*. 2012; 18(10): 49–53.
[32] Ozsolak F and Milos PM: RNA sequencing: Advances, challenges and opportunities. *Nat Rev Genet.* 2011; 12(2): 87–98.
[33] Wang L, Support Vector Machines: Theory and Applications, USA: Springer STUDFUZZ, 2005;177.
[34] Zhou Z-H and Feng J., Deep Forest: towards an alternative to deep neural networks, In: ArXiv e-prints 1702.08835v1, 2017.