Iraqi Journal of Information and Communication Technology(IJICT)
Vol. 5, Issue 2, August 2022
www.ijict.edu.iq
ISSN:2222-758X
e-ISSN: 2789-7362

LUNG CANCER RELAPSE PREDICTION USING PARALLEL XGBOOST

Rana D. Abdu-Aljabar [1], **Osama A. Awad** [2]
[1,2] College of Engineering, Al-Nahrain University, Baghdad, Iraq
ranadhiaa1@nahrainuniv.edu.iq [1], usamaawa@coie-nahrain.edu.iq [2]
Received:11/8/2021, Accepted:30/10/2021
DOI:10.31987/ijict.5.2.194

***Abstract*- Lung cancer has been the most popular form of cancer for decades. Surgery will offer the non-small cell lung cancer (NSCLC) patients the best hope of a cure if the cancer is diagnosed in the early stage. However, many patients eventually die of their disease due to relapse after surgery. Because of no symptoms of lung cancer in its early stage, many researchers try to improve methods to predict lung cancer relapse early. This study proposed a method to predict lung cancer relapse more accurately. This method has three stages; feature selection, parallel eXtreme Gradient Boost (XGBoost) classifications with different hyperparameters, and selection stage. It used two datasets of a gene expression microarray for different lung cancer types with its clinical information. The accuracy results of the proposed model are 0.88 and 0.83 for both datasets, which are more accurate than the represented machine learning. This multi-construction of the parallel XGBoost gives the system the flexibility to deal with a broader range of datasets without hyperparameters tuning and within a short time.**

keywords: **Machine learning, XGBoost, Lung cancer, Classification, Bioinformatics, Gene expression.**

I. INTRODUCTION

The term "lung cancer relapse" or "recurrence" refers to lung cancer disease that comes back after treatment. A relapse may be an alike or different type of previous cancer and may occur in the same or different location as before. Even with early-stage cancers and new cancer treatments, lung cancer recurrence happens rapidly, maybe in three months or more often than one might assume [1]. Most lung cancers recur in two to five years of the first diagnosis, depending on the cancer type and stage. The relapse rate in stage 1 of None Small Cell Lung Cancer (NSCLC) type is around three in 10 people, increasing to nearly seven in 10 by stage 4 [2]. Generally, early-stage tumor prediction has better clinical outcomes, and tumor staging aids treatment arranging. However, there are cases where patients unexpectedly produce recurrent disease, exemplifying the limitations of current clinical staging techniques in precisely predicting tumor recurrence. Many studies try to improve a method to predict lung cancer relapse early using gene expression profiles. They used different methods and had a good result, like Russul A. et al. [3]-[8]. She proposed different studies of new optimization models to improve NSCLC detection using microarray gene expression datasets. Also, Hasseeb A. et al.[9]-[12] have improved multiclass using Gene Expression Programming (GEP) algorithm in the lung cancer classification stage to determine specific therapy and reduce the fatality rate. Zhijun W. et al. [13] suggested a framework called DeepLRHE(a Deep convolutional neural network framework for Lung cancer recurrence Risk from Histopathology images Evaluation). It works on predicting the lung cancer recurrence risk by analyzing histopathological images of patients. Also, Shulong Li et al. [14] proposed a fusion algorithm that incorporates Handcrafted Features (HF) into the features learned in a 3D deep convolutional neural network's output layer. Patra R. [15] analyzed various machine learning classifier techniques to classify lung cancer into benign and malignant. Lai, Y et al. [16] trained clinical and gene expression data with an 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

ISSN:2222-758X
e-ISSN: 2789-7362

Mary Adline Priya [17] proposed an automatic approach to classifying the lung image into a normal case or cancer case by extracting noise from the CT lung file. The histogram analysis is then paired with morphological analysis, and lung regions are derived using thresholding operations. While the study of Adeola O. [18] used a clinical database to classify the patient if he has chronic kidney disease or not using XGBoost. In a previous study [19], we compared 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 applying a Parallel XGBoost (PXGB) with different hyperparameters to increase the system variety and decrease the overfitting. The PXGB showed more accurate prediction values for relapse and no relapse state.

II. XGBoost Algorithm

XGBoost is a decision-tree-based ensemble machine learning algorithm; it uses the Gradient Boosting approach to achieve machine learning algorithms, see Fig. 1. Tianqi Chen and Carlos Guestrin developed it. They introduced their work at the SIGKDD conference in 2016 [20]. It provides a parallel tree boosting that quickly and accurately solves many data science problems. In addition, it offers a range of hyperparameters that give fine-grained control over the model training procedure.

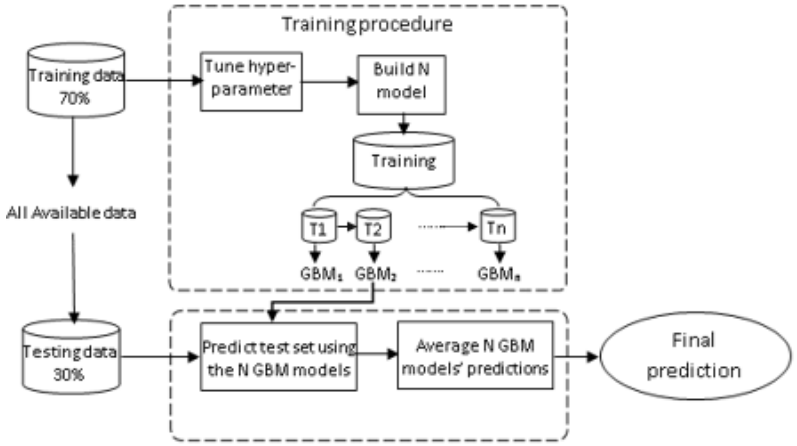

Figure 1: XGBoost model structure [21]

III. Lung Cancer Datasets

The datasets used in this study are microarray files. The data gathered through microarrays represents the gene expression profiles, which display changes in the expression of several genes simultaneously in response to a given disease or therapy. Thus, they represent the molecular level states of the cell [6]. This study applied the proposed model on two microarray datasets. Both datasets were downloaded from the National Center for Biotechnology Information site (NCBI).

ISSN:2222-758X
e-ISSN: 2789-7362

A. *Dataset Information*

This study used two gene expression microarray datasets with clinical information. The first one (GSE8894) dataset is an NSCLC type for 138 cases; 3 cases have no complete clinical information, so it becomes 135; 67 cases have lung cancer relapse state, and 68 cases have non-relapse lung cancer cases with their clinical information [22]. The second is (GSE68465) dataset. It also has clinical information and gene expression for 442 cases; after removing the incomplete information cases, 362 cases remain; 205 relapse cases and 157 non-relapse cases [14].

B. *Data Pre-Processing*

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 an irrelevant case and cause inappropriate learning results. The XGBoost classification deals with the numeric representation in the decision class. In contrast, classes in the lung cancer datasets are in nominal representation, like non-relapse / relapse. Therefore, it must change them to numeric representation (0 /1).

IV. THE PROPOSED METHOD

Decision tree-based algorithms are preferred for small to medium-sized structured/tabular files [19]. 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 XGBoosts structure to accommodate a broader type of datasets without changing its hyperparameters tuning. This method will be called the PXGB. It has three stages; the Feature selection stage, the parallel XGboost stage, and the selection stage, as shown in Fig. 2.

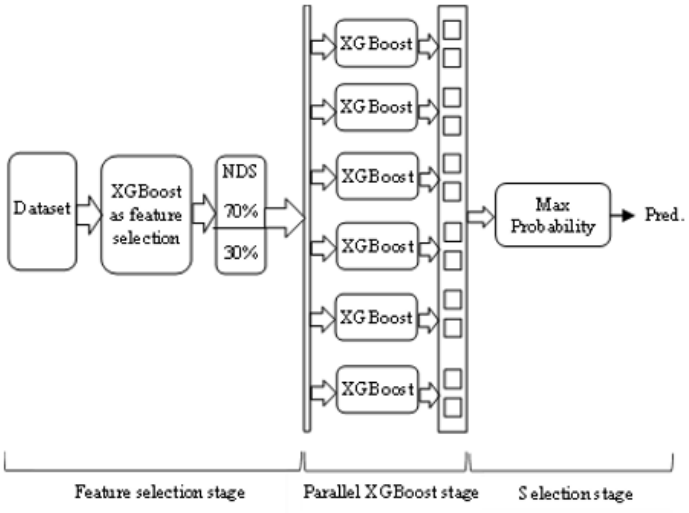

Figure 2: Block diagram of the proposed learning model (PXGB)

The feature selection stage: used the XGBoost module to rank the feature importance in making the prediction. In XGBoost, after constructing the boosted trees, each feature will be calculated its importance depending on how valuable

ISSN:2222-758X
e-ISSN: 2789-7362

the feature was in boosted trees construction. Each time the feature is used in the construction, the higher the importance score has. Thus, the importance score refers to how this feature is valuable or helpful in constructing the trees. The importance score algorithm calculated the importance for each decision tree by counting each feature cause splitting point and improving the performance measure, weighted by the number of observations for which the node is responsible. The feature's importance is then average across all decision trees within the model [17]. In this paper, the importance score threshold setting was all features above zero to take all features affecting tree construction. Each attribute less than this threshold will be neglected because it has no importance score. The GSE68219 features before the selection stage were 22283 features, and after it became 356 features, and so as for the GSE8894 dataset, the features were 54675, but after selection, they became 114 features. The parallel XGBoost stage: After the feature selection stage, New Dataset (NDS) will be used, which has only the effective features. The NDS will be split to 70% for training data and 30% for testing data, and then the training data will be entered into each XGBoost simultaneously. Each XGBoost has its hyperparameter set different from others (shown in Table I); these hyperparameter sets ranged from the most common values that may cause the overfitting to the most common values that may cause the underfitting. This kind of choice leads to having different XGBoost structures to be flexible to deal with various cases and datasets. All the XGboosts are working in parallel to reduce the extra overhead delays in learning time. Then the testing data will be applied to all XGBoost simultaneously to have different probability predictions for lung cancer relapse and no-relapse classes for each XGBoost model.

TABLE I
The setting of each XGBoost hyperparameters in the proposed model

XGBoost in the parallel stage	XGBoost hyperparameters				
	Sub-sample (ratio of the sample)	Max_ depth (tree level no.)	Learning rate	n_estimators (no of tree)	min_child_weight (no. of the sample)
First one	0.5	2	0.3	5	6
Second one	0.6	3	0.25	10	5
Third one	0.7	4	0.2	20	4
Fourth one	0.8	5	0.15	30	3
Fifth one	0.9	6	0.1	40	2
sixth one	1	7	0.05	50	1

Selection stage: This stage will take the maximum probability class of all XGBoost levels, considered the final class prediction.

V. THE RESULTS

The PXGB compared its results with original XGBoost, 2016 [20], Support Vector Machine (SVM) [23], the deep forest; multi-grained scanning (gcforest) [24], KNN (k-nearest neighbors algorithm), and Naive Bayes.

A. XGBoost Hyperparameters Setting

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

ISSN:2222-758X
e-ISSN: 2789-7362

TABLE II
Hyperparameter setting of representative models

XGBoost		SVM		gcForest		KNN		Naive Bayes	
hyperparameter	value	hyperparameter	value	hyperparameter	value	hyperparameter	value	hyperparameter	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_weight	1	C	1	Step	100	leaf_size	1		
subsample	0.7			min_samples_split	0.7				

B. The Comparison of different Classifiers

Applying the PXGB and other machine learning models to the lung cancer datasets has different prediction results for lung cancer relapse probability. The prediction values have four metric types:

TP: True Positive, which is in this study the correct prediction of lung cancer relapse.

TN: True Negative, which means the correct prediction of lung cancer no-relapse.

FP: False Positive, which means a false prediction of lung cancer relapse, while it is a no-relapse case.

FN: False Negative, which means a false prediction of lung cancer no relapse, while it is a relapse case.

The metrics used in this research for comparison and analyzing the efficiency of machine learning models are:

- Sensitivity: the true positive rate. Also called a recall

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{1}$$

- Specificity: the true negative rate.

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

- Precision is the truly detected lung cancer relapse divided by true and false detection of lung cancer relapse.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- F1-score : is a harmonic mean of precision and sensitivity. It can be used as a measure of performance of the test for the positive class.

$$F1 - score = 2 \cdot \frac{Sensitivity \cdot precision}{Sensitivity + precision} \tag{4}$$

- ROC: Receiver Operating Characteristic curve. It is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters, the True Positive Rate and the False Positive Rate.
- AUC: the Area Under Curve of the ROC.
- Accuracy is the proportion of all true predictions (lung cancer relapse or no-relapse cases) to all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

- Standard deviation: is a measure of variation between values in a set of data. The lower the standard deviation, the closer the data points to the mean or expected value; Conversely, a higher standard deviation indicates a broader range

of values.

$$Standarddeviation = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad (6)$$

x_i = The current (individual) prediction

μ= The mean of the predictions

N = number of the predictions

- learning time: The time spent in the Learning Stage (LS)

  learning time = The current time at the ending of the LS - The current time at the beginning of the LS

Table III illustrates each model's sensitivity, specificity, Precision, F1-score, AUC, accuracy, standard deviation, and learning time metrics. Furthermore, Figs. 3 and 4 show the ROC drawings and the AUC values of each machine learning model used in this study.

TABLE III
Comparison results of lung cancer relapse prediction

GSE8894 dataset								
Classifier Type	Sensitivity	Specificity	Precision	F1_score	AUC	Accuracy	Standard deviation	Learning time (min.)
PXGBS	0.85	0.90	0.89	0.87	0.88	0.88	0.058	00:09
XGBoost	0.55	0.67	0.61	0.58	0.61	0.61	0.061	00:10
SVM	0.65	0.43	0.52	0.58	0.54	0.54	0.061	00:04
gcForest	0.87	0.6	0.74	0.8	0.72	0.75	0.064	02:46
KNN	0.3	0.67	0.46	0.36	0.48	0.49	0.0733	00:10
Naive Bayes	0.75	0.62	0.65	0.7	0.68	0.68	0.0182	00:01
GSE68465 dataset								
PXGBS	0.95	0.66	0.79	0.86	0.81	0.83	0.011	00:16
XGBoost	0.81	0.57	0.71	0.76	0.69	0.71	0.0264	00:17
SVM	1.0	0.02	0.57	0.73	0.51	0.58	0.0280	00:11
gcForest	0.87	0.6	0.74	0.8	0.72	0.75	0.0425	03:01

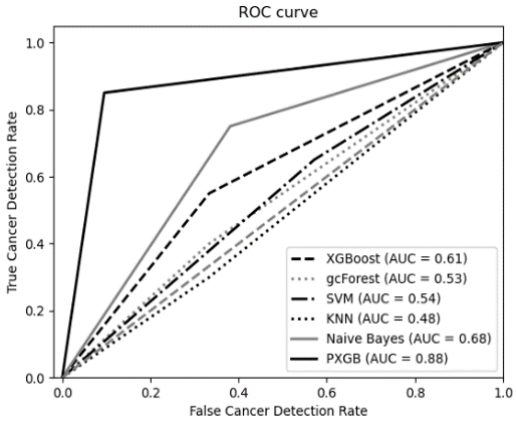

Figure 3: The ROC curves and AUC values for the GSE8894 dataset

ISSN:2222-758X
e-ISSN: 2789-7362

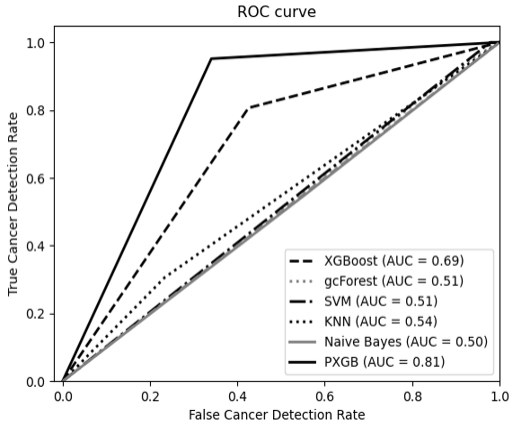

Figure 4: The ROC curves and their AUC values for the GSE68465 dataset

C. Analyzing Metrics

This study finds the sensitivity, specificity, Precision, F1-score, AUC, ROC, accuracy, standard deviation, and learning time metrics for each machine learning model used in this study to evaluate their effectiveness. When compared between the PXGB and the original XGBoost all its metric is better than it. That is because the PXGB depends on its prediction on different XGBoost structures that led to different probabilities in each case. Then in the selection stage, it chooses the maximum probability, making this algorithm a more accurate prediction than the original one, flexible in dealing with different datasets. The learning times in PXGB are 9 and 16 seconds for each dataset which is better than the original XGBoost (10, 17 seconds, respectively). That is because of the feature selection stage, which minimizes the number of features that cause speed up the learning stage, and although the PXGB has multiple XGBoost, they run in parallel, which leads to not consuming an extra overload to the learning time. Now it will analyze the results of PXGB with other machine learning. The PXGB results have the highest metric values among other machine learning when applied to the GSE8894 and GSE68465 datasets (as shown in Table III), except for the learning time value. The Naive Bayes has completed the learning stage in 1 second for both datasets, while PXGB completed it in 10 seconds to GSE8894 and 17 seconds to GSE68465, but it is still an acceptable value. The standard deviation values represented in Table III are taken from five runs of all machine learnings with different data in the learning stage at each run. It can be noticed that the PXGB's standard deviation value is less than most machine learnings, which indicates that the PXGB model is reliable even when dealing with different data. Furthermore, the results are illustrated in Figs. 5, 6, 7, 8, 9, 10, 11, and 12 as a histogram form.

ISSN:2222-758X
e-ISSN: 2789-7362

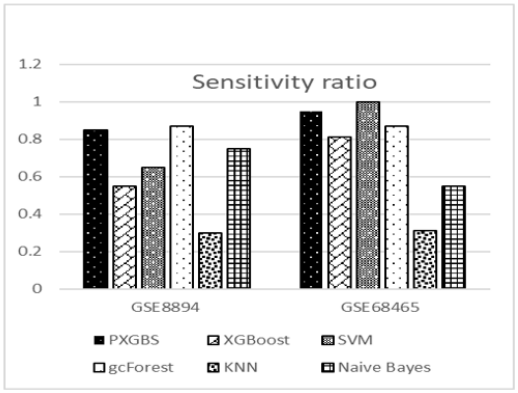

Figure 5: The sensitivity ratio chart

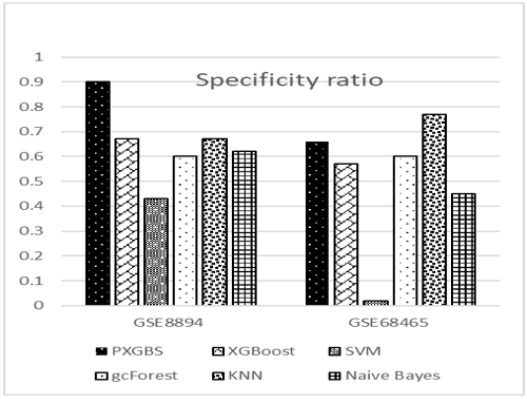

Figure 6: The specificity ratio chart

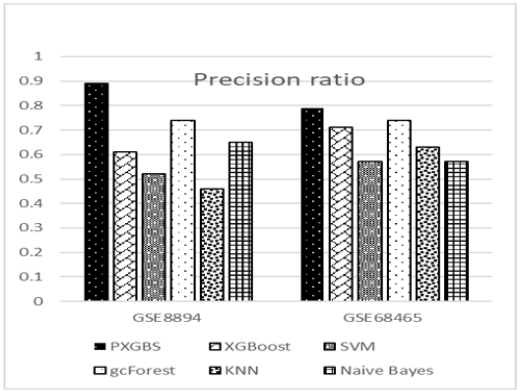

Figure 7: The precision ratio chart

ISSN:2222-758X
e-ISSN: 2789-7362

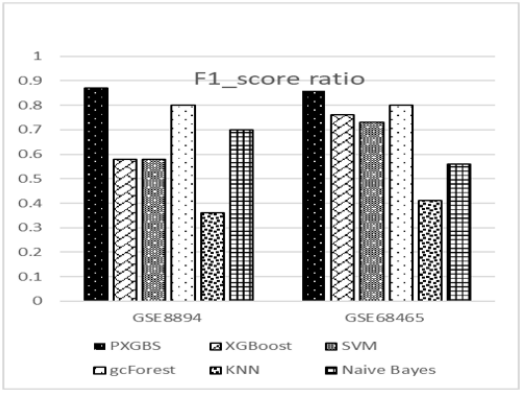

Figure 8: The F1-score ratio chart

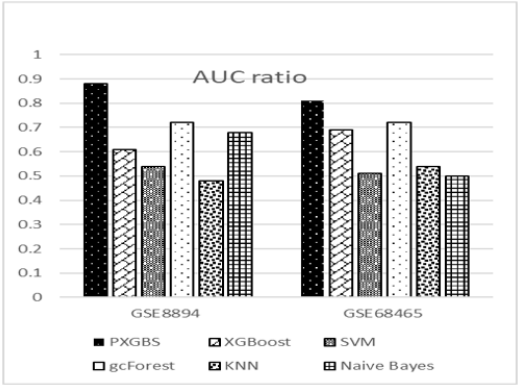

Figure 9: The AUC ratio chart

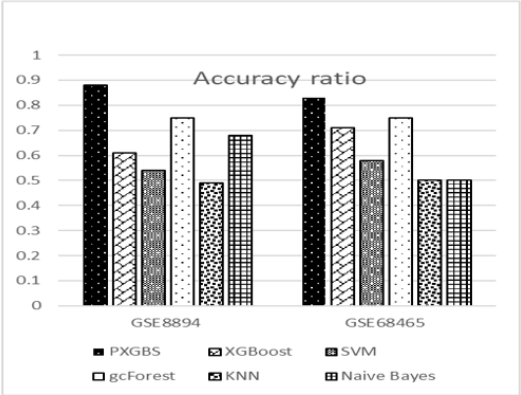

Figure 10: The accuracy ratio chart

ISSN:2222-758X
e-ISSN: 2789-7362

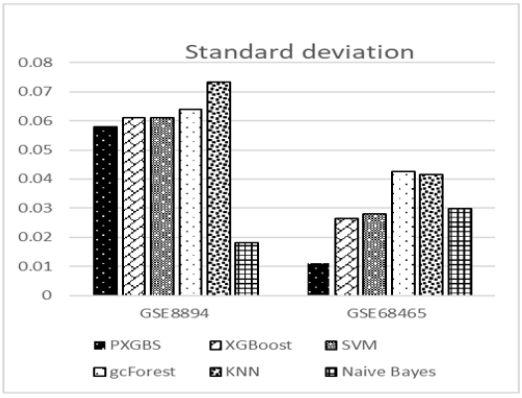

Figure 11: The standard deviation chart

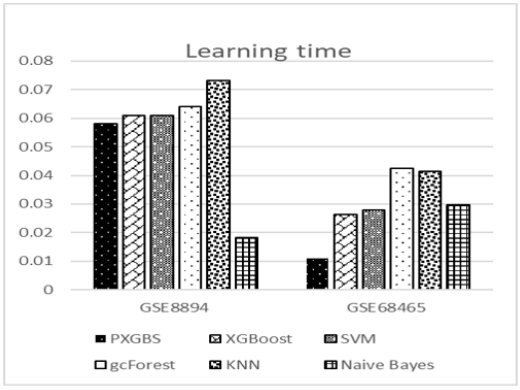

Figure 12: The learning time chart

VI. Conclusion

This study proposed a flexible prediction model using multi numbers of the XGBoost classifications connected in parallel. Each one has different hyperparameters to obtain various tree buildings. This variance in hyperparameters setting makes one or more of the XGBoosts perfect for a wide range of datasets when applied. That gives the model the flexibility to be applied to different datasets and has a good prediction. Using the XGBoost algorithm as a feature selection lets only active features associated with the learning process. That led to improving the accuracy and speeding up the learning time. The results showed that the PXGB achieved better accuracy prediction than other comparative machine learning. It also showed it provides accuracy more than the original XGBoost because it depends on building multi XGBoost structures in its learning stage, allowing it to deal with different datasets without tuning the hyperparameters by letting them have different probability values and choosing the highest one. Moreover, the system learned within a shorter time than the original. Furthermore, its small standard deviation value means it has a stable, reliable accuracy even when dealing with

different datasets. The result shows that the PXGB is flexible, reliable, and needs only a short time to deal with different datasets.

REFERENCES

[1] Choi PJ, Jeong SS, Yoon SS, "Prognosis of Recurrence After Complete Resection in Early-Stage Non-Small Cell Lung Cancer" , Korean J Thorac Cardiovasc Surg, Vol. 46, No. 6, pp. 449-456, 2013.
[2] Sasaki H, Suzuki A, Tatematsu T, et al, "Prognosis of Recurrent Non-Small Cell Lung Cancer Following Complete Resection" , Oncol Lett. , Vol. 7, No. 4, pp. 1300-1304, 2014.
[3] Al-Anni, R, Hou, J, Abdu-aljabar, R. D, et al, "Prediction of NSCLC Recurrence from Microarray Data with GEP" , IET Systems Biology, Vol. 11, No. 3, pp. 77-85, 2017.
[4] Al-Anni, R., Hou, J, Azzawi, H, et al, "Cancer Adjuvant Chemotherapy Prediction Model for Non-Small Cell Lung Cancer" , IET Systems Biology, Vol. 13, No. 3, 2018.
[5] Al-Anni, R, Hou, J, Azzawi, H, et al, "Risk Classification for NSCLC Survival Using Microarray and Clinical Data" , Proc. of 207th The IIER Int. Conf. , 12th-13th December, Paris, France, 2018.
[6] Al-Anni, R, Hou J, Azzawi, H, et al, "A Novel Gene Selection Algorithm for Cancer Classification Using Microarray Datasets" , BMC Med. Genomics, Vol. 12, No. 10, 2018.
[7] Al-Anni, R, Hou, J, Azzawi, H et al, "Deep Gene Selection Method to Select Genes from Microarray Datasets for Cancer Classification" , BMC-Informatics, Vol. 20, No. 608, 2019.
[8] Al-Anni, R, Hou, J, Azzawi, H et al, "New Gene Selection Method Using Gene Expression Programming Approach on Microarray Data Sets" , Lee R. (eds) Computer and Information Science, Studies in Computational Intelligence, Springer, Vol. 791, pp. 17-31, 2018.
[9] Azzawi, H, Hou, J, Xiang, Y. et al, "Lung Cancer Prediction from Microarray Data by Gene Expression Programming" , IET Syst. Biol. , Vol. 10, No. 5, pp. 168-178, 2016.
[10] Azzawi, H, Hou, J, Alanni, R, et al, "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 38, pp. 541-553, 2017.
[11] Azzawi, H, Hou, J, Alanni, R, et al, "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, Singapore IEEE, 6-8 June,pp. 68-73, 2018.
[12] Azzawi, H., Hou, J, Alanni, R, et al, "A Hybrid Neural Network Approach for Lung Cancer Classification with Gene Expression Dataset and Prior Biological Knowledge" , Int. Conf. on Machine Learning for Networking, Paris, France, Springer, Cham, Lecture Notes in Computer Science, Vol. 11407, pp. 279-293, 2019.
[13] Wu, Z, Wang, L, Li, C, et al, "DeepLRHE: A Deep Convolutional Neural Network Framework to Evaluate the Risk of Lung Cancer Recurrence and Metastasis From Histopathology Images" , Frontier in Genetic, Vol. 11, No. 768, 2020.
[14] Li, S, Xu, P, Li, B, el al, "Predicting Lung Nodule Malignancies by Combining Deep Convolutional Neural Network and Handcrafted Features" , Physics in Medicine & Biology, Vol. 64, No.17, 2019.
[15] Patra, R, "Prediction of Lung Cancer Using Machine Learning Classifier" , Int. Conf. on Computing Science, Communication and Security Computing Science, Communication and Security, Springer, Singapore, Vol. 1235, pp. 132-142, 2020.
[16] Lai, Y, Chen, W, Hsu, T, et al, "Overall Survival Prediction of Non-Small Cell Lung Cancer by Integrating Microarray and Clinical Data with Deep Learning" , Sci. Rep. Nature research, Vol. 10, No. 4679, 2020.
[17] Priya, M. M. A, Jawhar, S. J, "Advanced Lung Cancer Classification Approach Adopting Modified Graph Clustering and Whale Optimisation-Based Feature Selection Technique Accompanied by A Hybrid Ensemble Classifier" , IET, Vol. 14, No. 10, pp. 2204-2215, 2020.
[18] Ogunleye A, Wang Q-G, "XGBoost Model for Chronic Kidney Disease Diagnosis" , IEEE/ACM, Trans Comput Biol Bioinform, Vol. 17, No. 6, pp. 2131-2140, 2020.
[19] Abdu-aljabar, R. D, Awad, O. A, "A Comparative Analysis Study of Lung Cancer Detection and Relapse Prediction Using XGBoost Classifier" , 2nd Int. Sci. Conf. of Eng. Sci. (ISCES2020), Col. of Eng. , Univ. of Diyala & IOP Pub. , Vol. 1076, 2021.
[20] Chen, T, Guestrin, C, "XGBoost: A Scalable Tree Boosting System" , In Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 785-794, 2016.
[21] Yuanyuan Li, David M. Umbach, Adrienna Bingham, Qi-Jing Li, Yuan Zhuang and Leping Li1, "Putative Biomarkers for Predicting Tumor Sample Purity Based on Gene Expression Data" , BMC Genomics, Vol. 20, No. 1021, 2019.
[22] Mezheyeuski, A, Bergsland C. H, Backman, M, 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.
[23] Wang, L, "Support Vector Machines: Theory and Applications" , USA: Springer, 2005.
[24] Zhou, Z-H, Feng, J, "Deep Forest: Towards An Alternative to Deep Neural Networks" , Proc. of the 26th Int. Joint Conf. on AI, IJCAI-17, pp. 3553-3559, 2017.

