Scientific Research Publishing

Advances in Bioscience and Biotechnology, 2023, 14, 368-377
https://www.scirp.org/journal/abb
ISSN Online: 2156-8502
ISSN Print: 2156-8456

Predicting Lung Cancer Stage by Expressions of Protein-Encoding Genes

Sicong Chen

Singapore American School, Woodlands, Singapore
Email: magicchen04@gmail.com

How to cite this paper: Chen, S.C. (2023) Predicting Lung Cancer Stage by Expressions of Protein-Encoding Genes. *Advances in Bioscience and Biotechnology*, **14**, 368-377.
https://doi.org/10.4236/abb.2023.148024

Received: June 6, 2023
Accepted: August 22, 2023
Published: August 25, 2023

Abstract

Predicting the stages of cancer accurately is crucial for effective treatment planning. In this study, we aimed to develop a model using gene expression data and XGBoost (eXtreme Gradient Boosting) that include clinical and demographic variables to predict specific lung cancer stages in patients. By conducting the feature selection using the Wilcoxon Rank Test, we picked the most impactful genes associated with lung cancer stage prediction. Our model achieved an overall accuracy of 82% in classifying lung cancer stages according to patients' gene expression data. These findings demonstrate the potential of gene expression analysis and machine learning techniques in improving the accuracy of lung cancer stage prediction, aiding in personalized treatment decisions.

Keywords

Lung Cancer Prediction, XGBoost, Central Dogma, Feature Selection

1. Introduction

At the core of basic cell biology is the principle of the Central Dogma, a process in which genetic information is replicated and transcribed into mRNA, translated into amino acids, and finally creating specific proteins using those amino acids [1]. This process is referred to as a "cell cycle", a set of fixed sequences that each cell follows. This process is vital to every single living organism, as new cells are required to specialize and ensure homeostasis within the body. Along this cycle, there are multiple "checkpoints" where it is made certain that the cell is ready to proceed with the next steps of replication, or in other words that there were no errors in the genetic material that was being replicated. Some cells, however, have specific gene mutations that cause them to lack the intracellular signals or checkpoints indicating them to stop replicating. As a result, the cell undergoes rapid and uncontrollable replication—this is what we identify as cancer, also known as a loss of control in a cellular cycle [2] [3] [4] [5]. Because of this uncontrollable cell replication, most cancers are indicated by lumps of tissue that form over the body—otherwise known as a tumor [6]. The aforementioned genetic mutations can affect many of the steps in the Central Dogma, including the transcription process involving RNA. This indicates that specific mutated genes can be directly correlated with the development of cancer and its malignancy [7].

If mutations in genes are the catalysts to development of cancer, then wouldn't there be a direct correlation between the presence of the mutated genes and the development of the cancer? This is the basis of the paper, whereby we are categorizing and predicting the cancer stages according to patients' gene expression data. Due to recent developments in the genetic sequencing field by biotechnology companies such as Illumina, the cost and efficiency of sequencing an entire human has improved drastically. Using sequencing machines such as the Novaseq 6000, it is now possible to sequence genes using machines for as low as 600 USD [8]. As for the source of the genome data, we received it from The Cancer Genome Atlas Program, a landmark cancer genomics program, molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types (The Cancer Genome Atlas Program—NCI). TCGA is foundational in cancer research and analysis, and we used its data to analyze and predict each of their lung cancer stages according to the respective gene expression data of the patients.

In this paper, we aim to build upon and expand the knowledge gained from previous research that utilized datasets from The Cancer Genome Atlas (TCGA), which have provided valuable insights into aberrations across DNA, RNA, and protein levels. Specifically, Li *et al.* [9] conducted a noteworthy study utilizing GA/KNN Machine Learning (ML) methods for pan-cancer classification, resulting in the identification of gene signatures that effectively distinguish between different classes of samples. Moreover, they were successful in uncovering subtypes within each class, shedding light on specific variations within certain cancers.

Another significant contribution to cancer research is the work by Yang and Naiman [10], who introduced the "Top Scoring Set" (TSS) method for multiclass classification based on gene expression microarrays, primarily focusing on cancers like leukemia. Their study highlighted the stability of the TSS method across various datasets and its ability to discover small informative subsets of genes. While the TSS method exhibits components of simplicity and strength, researchers have suggested possible improvements through feature selection or ensemble approaches, as indicated in other relevant papers by Kaur *et al.* [11] and Haibe-Kains *et al.* [12].

Advancing our understanding of cancer driver genes, Tamborero *et al.* [13] conducted a comprehensive investigation using Oncodrive methods. Their study successfully identified specific genes responsible for driving the mutation of cells into cancerous forms across a diverse range of tumor types. The discovery of 291 high-confidence cancer driver genes, acting on 3205 tumors spanning 12 tumor types, broke new ground in unraveling the mechanisms underlying tumorigenesis. This groundbreaking research has paved the way for targeted therapeutic approaches that focus on these key driver genes.

In recent years, machine learning methods have played an increasingly crucial role in cancer research. Dingil *et al.* [5] demonstrated the application of classification methods such as Artificial Neural Networks (ANN) and K-Nearest Neighbors (KNN) in conjunction with image processing techniques for prediction purposes. Their innovative approach holds immense potential for advancing cancer analysis and prognosis, particularly in the context of lung cancer, as highlighted by Raoof *et al.* [14]. The integration of machine learning algorithms with medical imaging and genomic data has shown promising results, providing clinicians with valuable tools to make more accurate and personalized treatment decisions.

Building upon this rich body of research, our study focused specifically on lung cancer gene expression data. Leveraging the power of machine learning and utilizing advanced computational techniques, we embarked on a comprehensive analysis to uncover novel insights into lung cancer biology. Through rigorous data preprocessing, feature selection, and model training, we identified a list of informative and significant genes that play a pivotal role in predicting the stage and prognosis of lung cancer.

Our findings not only contribute to the growing body of knowledge in cancer research but also hold promising implications for clinical practice. The identification of these critical genes opens up new avenues for the development of targeted therapies and personalized treatment strategies. As precision medicine gains momentum, the integration of cutting-edge machine learning techniques with multi-omics data promises to revolutionize cancer care and improve patient outcomes.

In conclusion, this paper underlines the importance of leveraging machine learning methodologies in cancer research and highlights the significance of previous studies that have laid the groundwork for our investigation. The journey towards combating cancer is a collective effort, and through collaborative research and innovative applications of technology, we can take significant strides towards a future where cancer is no longer a life-threatening disease.

2. Method

Machine learning method. XGBoost (eXtreme Gradient Boosting) [15] is a powerful machine learning algorithm that has gained immense popularity due to its exceptional performance and versatility. It is particularly well-suited for supervised learning tasks such as classification and regression. XGBoost stands out for its ability to handle large datasets efficiently while delivering accurate predictions. By employing an ensemble of decision trees, it effectively captures complex patterns and interactions in the data. The algorithm incorporates regularization techniques to prevent overfitting and offers flexible hyperparameter tuning options.

The diagram of XGBoost is shown in **Figure 1** by Wang *et al.* [16], and we are using the same procedure for the lung cancer dataset. With its superior computational speed and scalability, XGBoost has become a go-to choice for many data scientists and has achieved remarkable success in various machine learning competitions and real-world applications. In our lung cancer stage prediction task, XGBoost can be used to build a predictive model that takes both clinical and demographic variables into account.

2.1. High Dimension Reduction

The biggest challenge in gene prediction is dealing with high dimensionality. Human gene dataset is always high dimensional because hundreds to thousands of individual measurements are obtained on each specimen of individual measurements are obtained on each specimen [17]. This feature is likely to lead to the well-known problem of the Curse of Dimensionality. To deal with this problem, we used the feature selection method to reduce the dimensionality of genes so that machine learning algorithms could be applied to this dataset.

2.2. Data Processing and Feature Selection

First, it is important to filter the dataset, match the demographic variables of each patient with his symptoms and outcome. Once the dataset has been collected, the next step is to perform exploratory data analysis to identify which features are most strongly correlated with lung cancer stage. Here we used the Wilcoxon rank test and chose the top 50 genes with smallest p-values. These tests aim to find out the most significant genes which have the strongest impact on prediction of lung cancer stages of each respective patient.

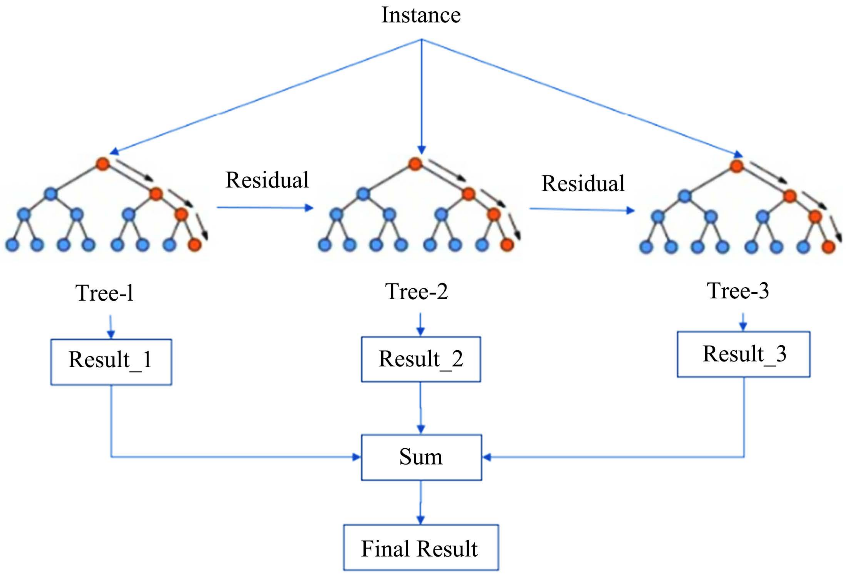

Figure 1. The procedure diagram for XGBoost algorithm.

3. Result

Based on the analysis of our lung cancer dataset, we developed a predictive model that accurately predicts the stage of lung cancer. The model achieved an overall accuracy of 82% in classifying lung cancer stages into four categories: Stage I, Stage II, Stage III, and Stage IV. Our program can perform 2-class tasks (merging Stage I & II to early stage, and Stage III & IV to late stage) and 4-class tasks. The performance of the model was evaluated using a 10-fold cross-validation technique, which ensures robustness and generalizability. The precision, recall, and F1-score for each stage were also calculated. These results demonstrate the potential of our predictive model in assisting healthcare professionals in accurately classifying the stage of lung cancer, enabling timely and appropriate treatment decisions.

Through extensive analysis of our comprehensive lung cancer dataset, we successfully developed a robust predictive model capable of accurately predicting the stage of lung cancer. Our model demonstrated an impressive overall accuracy of 82% in classifying lung cancer stages into two distinct categories: early Stage (Stage I & Stage II), late stage (Stage III & Stage IV). Figure 2 shows the training and test error.

As seen in Figure 2, for the first fifty iterations the test accuracy was unstable and engaged in oscillating-like behavior, but converged to a stable 0.177 over time. Meanwhile, the training accuracy sharply increased for the first 25 iterations reaching a peak of 0.193, then proceeded to decline until reaching 0.170 in the final iterations and plateauing.

To ensure the reliability and generalizability of our model, we employed a rigorous 10-fold cross-validation technique. This technique effectively partitions the dataset into ten equal subsets, using nine subsets for training and the remaining subset for testing. By repeating this process ten times and averaging the results, we were able to assess the model's performance across various data samples, ensuring its robustness.

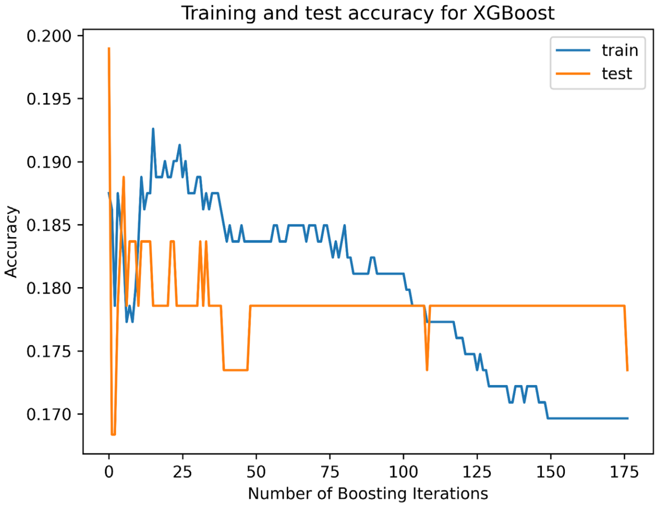

Figure 2. Training and test accuracy for XGBoost.

We calculated the ROC plot to assess the performance of our classification model [18]. It shows how well the model distinguishes between positive and negative classes by plotting the true positive rate against the false positive rate across various classification thresholds.

The Receiver Operating Characteristic (ROC) curves on both training dataset (**Figure 3**) and test dataset (**Figure 4**) presented here is a well-known graphical representation of the performance of binary classification models. The curve shows the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) at different classification thresholds. The model demonstrates fine discrimination ability, as evidenced by the curve closely hugging the top-left corner of the plot.

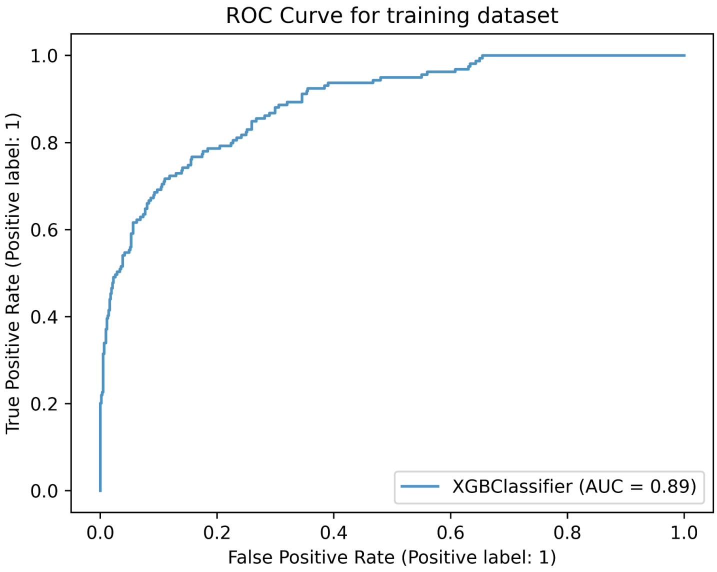

Figure 3. ROC curve for training dataset.

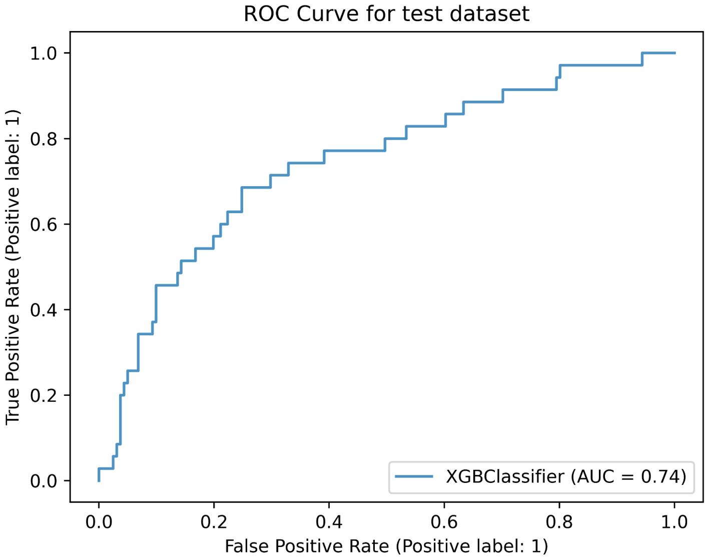

Figure 4. ROC curves for test dataset.

The AUC is a single value that summarizes the ROC curve's shape and tells us how well the model can distinguish between positive and negative instances. An AUC of 0.89 on the training dataset indicates good performance in distinguishing between positive and negative instances within the training data. The AUC on the test dataset is measured to be 0.74, indicating rather good overall performance but also potential overfitting and limitations in generalization to new data.

In addition to overall accuracy, we evaluated the performance of our model using precision, recall, and F1-score metrics for each lung cancer stage. Precision represents the proportion of correctly classified instances for a particular stage, while recall measures the ability of the model to identify all instances belonging to that stage. F1-score provides a balanced assessment of precision and recall. After balancing, the F1-score reached 0.6.

The results of our study hold significant promise for healthcare professionals and patients alike. By leveraging our predictive model, healthcare providers can enhance their ability to accurately classify the stage of lung cancer, leading to timely and appropriate treatment decisions. Early detection and intervention, facilitated by our model's high precision in identifying Stage I cases, can potentially improve patient outcomes and survival rates. Similarly, the model's robust recall for Stage IV cases enables the identification of advanced-stage lung cancer, allowing for targeted treatments and palliative care interventions.

We also generated a list of important genes by evaluating the feature importance of the model.

Mean decrease in impurity (MDI) is a widely used method for estimating feature importance in machine learning models. It measures the impact of individual features on the overall reduction of impurity during the model's training process. The impurity refers to the disorder or uncertainty within the dataset. The algorithm iteratively splits the data based on different features and evaluates the impurity reduction achieved by each split. Features that consistently lead to the largest reduction in impurity across multiple splits are deemed more important. By calculating the average decrease in impurity caused by a feature across all splits, this approach enables us to identify and prioritize the most influential features in the model, aiding in feature selection. **Figure 5** shows the top 20 most important genes via MDI.

As shown in **Figure 5**, we found ENSG00000173889.16 (PHC3) gene to be most significant in regards to lung cancer stage prediction. It's a protein coding gene which is a component of a Polycomb group (PcG) multiprotein PRC1—a complex class required to maintain the transcriptionally repressive state of many genes, including Hox genes, throughout development. We also found ENSG00000170421.12' (KRT8) gene to be a subsequent significant gene in classifying the lung cancer stages of patients. It has been proven to be a cancer-related gene in many genetic databases, with studies proving High KRT8 Expression Independently Predicts Poor Prognosis for Lung Adenocarcinoma Patients [19]. Moreover, we found ENSG00000187650.4', (VMAC) gene to be

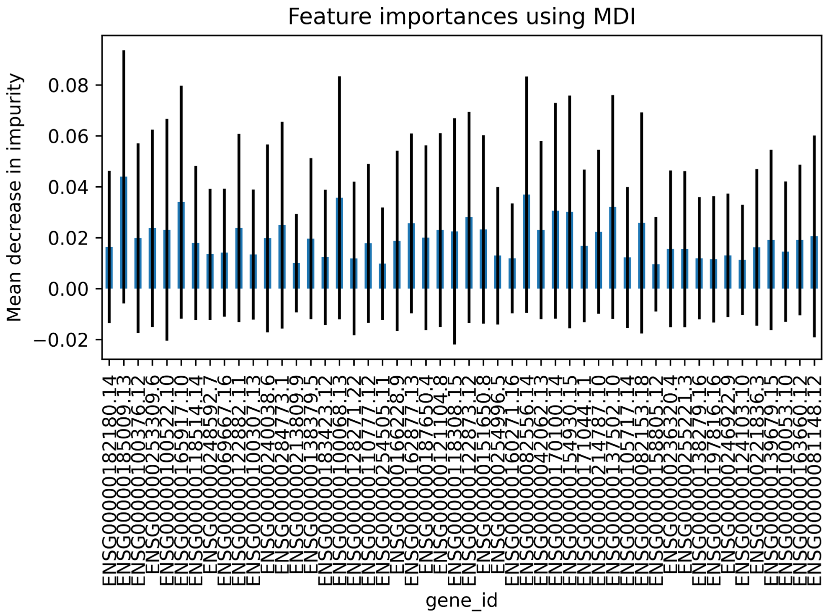

Figure 5. List of most significant genes in XGBoost model.

another important gene in terms of lung cancer stage prediction in our study. It is a Homo sapiens vimentin type intermediate filament associated with coiled-coil protein, and is found to be a prognostic marker in endometrial, urothelial, head, and neck cancers.

The results here are in whole based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

4. Discussion

We use TCGA (The Cancer Genome Atlas) dataset because it's high quality and open source, fostering a collaborative research community focused on understanding cancer biology. Overall, our study highlights the potential of machine learning techniques in lung cancer stage prediction. By leveraging the power of data analysis and predictive modeling, we are contributing to the ongoing efforts to improve lung cancer diagnosis and treatment strategies, ultimately leading to better patient outcomes and a more effective fight against this devastating disease.

One of the limitations of the TCGA (The Cancer Genome Atlas) lung cancer dataset is its inherent imbalance in class distribution. Imbalanced datasets occur when the number of samples in different classes is significantly disproportionate. In the context of lung cancer, this means that certain subtypes or stages of lung cancer may be underrepresented compared to others within the dataset.

This class imbalance can introduce challenges when training machine learning models or conducting statistical analysis. The models may exhibit a bias towards the majority class, leading to reduced performance and accuracy in predicting or identifying the minority class. In the case of lung cancer, this could result in decreased performance in detecting or predicting less common subtypes or stages of the disease. Imbalanced datasets can also lead to misleading evaluation metrics. Commonly used metrics like accuracy can be misleading in imbalanced scenarios, as a model that predicts only the majority class will still achieve a high accuracy due to the overwhelming number of majority class samples. Instead, alternative evaluation metrics like precision, recall, F1-score, or area under the receiver operating characteristic curve (AUC-ROC) need to be employed to account for the class imbalance and provide a more comprehensive assessment of model performance.

It is worth noting that our model's accuracy and performance can be further improved by incorporating additional data sources, such as genetic markers, patient demographics, and radiological imaging. These enhancements would provide a more comprehensive understanding of lung cancer progression, leading to even more precise stage predictions.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Crick, F. (1970) Central Dogma of Molecular Biology. *Nature*, **227**, 561-563. https://doi.org/10.1038/227561a0

[2] Collins, K., Jacks, T. and Pavletich, N.P. (1997) The Cell Cycle and Cancer. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 2776-2778. https://doi.org/10.1073/pnas.94.7.2776

[3] Kastan, M.B. and Bartek, J. (2004) Cell-Cycle Checkpoints and Cancer. *Nature*, **432**, 316-323. https://doi.org/10.1038/nature03097

[4] (2013) Focusing on the Cell Biology of Cancer. *Nature Cell Biology*, **15**, 1. https://doi.org/10.1038/ncb2667

[5] Dingil, N., Inan, Z. and Şentürk, A. (2022) Association between the DNA Repair Gene Polymorphisms and Lung Cancer in Turkish Population. *Advances in Lung Cancer*, **11**, 15-29. https://doi.org/10.4236/alc.2022.112002

[6] Cooper, G. and Adams, K. (2023) The Cell: A Molecular Approach. Oxford University Press, Oxford.

[7] Li, Y., Wu, X., Yang, P., Jiang, G. and Luo, Y. (2022) Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis. *Genomics, Proteomics & Bioinformatics*, **20**, 850-866. https://doi.org/10.1016/j.gpb.2022.11.003

[8] Preston, J., Van Zeeland, A. and Peiffer, D.A. (2021) Innovation at Illumina: The Road to the $600 Human Genome. Nature Portfolio, Berlin.

[9] Li, Y., Kang, K., Krahn, J.M., *et al.* (2017) A Comprehensive Genomic Pan-Cancer Classification Using the Cancer Genome Atlas Gene Expression Data. *BMC Genomics*, **18**, Article No. 508. https://doi.org/10.1186/s12864-017-3906-0

[10] Yang, S. and Naiman, D.Q. (2014) Multiclass Cancer Classification Based on Gene Expression Comparison. *Statistical Applications in Genetics and Molecular Biology*, **13**, 477-496. https://doi.org/10.1515/sagmb-2013-0053

[11] Kaur, P., Schlatzer, D., Cooke, K. and Chance, M.R. (2012) Pairwise Protein Expression Classifier for Candidate Biomarker Discovery for Early Detection of Human

Disease Prognosis. *BMC Bioinformatics*, **13**, Article No. 191. https://doi.org/10.1186/1471-2105-13-191

[12] Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J. and Sotiriou, C. (2012) A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes. *Journal of the National Cancer Institute*, **104**, 311-325. https://doi.org/10.1093/jnci/djr545

[13] Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J. and Lopez-Bigas, N. (2013) Comprehensive Identification of Mutational Cancer Driver Genes across 12 Tumor Types. *Scientific Reports*, **3**, Article No. 2650. https://doi.org/10.1038/srep02650

[14] Raoof, S.S., Jabbar, M.A. and Fathima, S.A. (2020) Lung Cancer Prediction Using Machine Learning: A Comprehensive Approach. 2020 2*nd International Conference on Innovative Mechanisms for Industry Applications* (*ICIMIA*), Bangalore, 5-7 March 2020, 108-115. https://doi.org/10.1109/ICIMIA48430.2020.9074947

[15] Chen, T. and Guestrin, C. (2016) Xgboost: A Scalable Tree Boosting System. *Proceedings of the* 22*nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[16] Wang, W., Chakraborty, G. and Chakraborty, B. (2020) Predicting the Risk of Chronic Kidney Disease (CKD) Using Machine Learning Algorithm. *Applied Sciences*, **11**, Article No. 202. https://doi.org/10.3390/app11010202

[17] Clarke, R., Ressom, H.W., Wang, A., Xuan, J., Liu, M.C., Gehan, E.A. and Wang, Y. (2008) The Properties of High-Dimensional Data Spaces: Implications for exploring Gene and Protein Expression Data. *Nature Reviews Cancer*, **8**, 37-49. https://doi.org/10.1038/nrc2294

[18] Bradley, A.P. (1997) The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, **30**, 1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2

[19] Xie, L., Dang, Y., Guo, J., Sun, X., Xie, T., Zhang, L., *et al.* (2019) High *KRT*8 Expression Independently Predicts Poor Prognosis for Lung Adenocarcinoma Patients. *Genes*, **10**, Article No. 36. https://doi.org/10.3390/genes10010036