JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

LUNG CANCER PREDICTION

Rahul Yadav
School of Computer Application.
Professional University,
India

Sachin Yadav
School of Computer Application
Lovely Professional University,
Punjab, India

Gaurav Singh
School of computer Application Lovely
Lovely Professional University,Punjab,
Punjab, India

Mr. Vivek Kumar Sharma
(Assistant professor, SCA) Lovely professional University

ABSTRACT

The lung cancer is considered to be the deadliest type of a disease. At the moment we cannot diagnose it in time without the involvement of the medical staff. Despite the fact that we are still far away from ultimate understanding of cancer's mechanisms and a solid cure, early diagnosis significantly improve the odds of successful treatment. By taking up the new technology including machine learning, image processing and many more, there would be a promising way for accurate diagnosis and prediction of cancer. In the latest experiments, scientists concentrated on designing an accurate method that could be valuable in image processing and machine learning to classify and predict lung cancer. The data gathering started with gathering images which was accomplished with 83 CT scans data from 70 different patients as the dataset. Before the segmentation, the images undergo preprocessing, such as noise reduction and enhancement through geometric mean filtering, thus, upgrading image quality. Additionally, the Khan is divided into healthy and affected areas using the cluster analysis method. This approach enabled targeting areas in which cancer had occurred, hence giving a good platform on which the identification and forecast of cancer would be carried out. Additionally, the ANN, KNN, and RF machine learning algorithms were utilized for the classification process. A comparison showed that the ANN model was the most consistent in their ability to predict lung cancer cases.

Keywords:
Data preprocessing, Data Visualization
Feature selection ,Model building, Model evaluation

1. INTRODUCTION

Lung cancer, one of the biggest killers of the society, has cheated every one million people again and again of their lives only this year. In the light of current medical situation, the lung nodule identification using the chest computed tomography scans turns out to be indispensable. The dramatic rise in the number of lung nodules relays the significance of the introduction of computerized detection (CAD) systems that would serve as the first line of defense against lung cancer [1] As mentioned in 2018, the disease of lung cancer might only be perceivable to have killed about 9.1 third of each cancer-related 6 million lives, becoming the leading cause of cancer death. deaths worldwide. Among cancers, lung cancer is often a rallying point for public dialogue. cancer types and their respective occurrance. Approximately 2.09 million lung cancers are projected worldwide, with deaths of 1.76 million people; this represents nearly 84% of all deaths from cancer-

related causes. Lung cancer is known as the deadliest disease to humankind for it owns a high rate of rapidly multiplying and becoming abnormal cells in the lungs forming tumors in the whole. These cancerous cells have not only the propensity to spread very fast but also the arteries and lymph streams can be the aides to make them spread exclusively within lung tissue. Most likely, the spread of cancer cells is enhanced by the natural movement of lymph-propelled cells tending to go to the mediastinum, the central part of the chest region (the air passageway). As the cancer cells proliferate and invade on the neighboring tissues and organs, the process called metastasis inevitably transpires [2]. During a CT scan, advanced X-ray equipment is employed to capture images of the human body from various angles. Subsequently, these images are input into a computer system, where they undergo processing to generate a cross-sectional depiction of the body's internal organs and tissues [3].

Air, as well, performs a similar function in our body, as it flows down through our nasal cavity, pharynx, larynx, and trachea to reach our lungs. The trachea splits as it gets inside the chest until it becomes any thinner than the original tube. It then finally divides into a tuft of smaller branches called alveoli which are mainly made of air sacs. In most cases the alveoli have plenty of capillaries that are their distinct factor. As a result there will be carbon dioxide extraction from the blood, and oxygen will be added. The breathing act remains the most joyful he is, no matter where he is situated on the earth, each of his breaths give oxygen for the blood to his lungs, which is important for his life [5].In many developed countries, lung diseases remain a primary cause for passing away. For example, the risk factors for instance, smoking, exposure to environmental pollutants, and chronic inflammation work together to inflict even worse damage, which is often beyond repair. Although the lungs have adapted to contain natural mechanisms of self-clearance (e.g., phlegm production), smoking typically overcomes these processes. Additionally, a complex system of environmental factors, genetics and heritance attributes also contributed to the diseases progression by affecting lung health. Diseases that affect the respiratory system come in variety of categories giving numerous difficulties and repercussions [5].

The inserting of the X-ray dye intravenously by injection is probably to do great for the quality of CT images so that you can see a lot more complicated organs and tissues. It thus becomes one of the several potential advantages of Contrast Injection CT imaging. CT scan shows the reliability of the abdominal or pelvic region such as the fluid accumulation or the enlarged lymph nodes in this region. Moreover, they can as well locate gallstones and kidney stones very well. CT scans might not be as specific about some organs than those imaging tests like PET or MRIs.

Whenever abnormalities are detected in liver or the soft tissues around it as stomach they Therefore offer an indirect assessment of these organs [5,6].

2. LITERATURE REVIEW

Cancer is still the major and most deadly health issue in the 21st century, which can be observed through the increasing cases of lung cancer, which is the cause of both morbidity and mortality. Lung cancer is generally classified into three major kinds: non-small cell lung cancer (NSCLC), which in turn is subdivided into lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD), that together make up around 85% of the cases of lung cancer [3]. Accurate subcategorization of lung cancer is very important for both the prognosis and the treatment plan which affect patients' survival the most.

Investigative noninvasive imaging methods such as positron emission tomography (PET) and computed tomography (CT) make their contribution to lung cancer diagnosis [8]. IHC Evaluation is the gold standard for classification, however, it requires invasive tissue biopsies which in turn, results in delayed diagnosis and suffering of the patients.

Developments in the field of artificial intelligence have enabled lung cancer diagnosis via the interpretation of data from CT and PET images, as well as tissue samples. Deep learning approaches have demonstrated their ability to discriminate nodules that are benign from malignant ones, Kaggle completion being a perfect illustration of this point [7,9]. Recent deep learning and radiomics techniques have enlarged beyond the benign/malignant classification, integrating the quantitative features from PET images which predicts metastasis cancer with 6and employing the deep learning techniques to pathological images for NSCLC and SCLC sub-types classification [10].studies

The variety of diagnostic techniques in thoracic cancer can be observed with different imaging modalities being implemented [11]. Deep learning models have proven their ability of blending of features from multiple modalities like PET/CT data for improved diagnosis [12].

Fine-grained classification of lung cancer types presents tasks which can be attributed to difficulty in acquiring relevant features in the small lesion regions and worrying about feature noise. The problem with channel-wise non-linearity has been countered by employing spatial and channel-wise attention mechanisms [12-13]. Nevertheless, the optimization of attention mechanisms for multimodality datasets is a field that needs continued development.

There are many approaches to predictive modeling, which take into account the methods like fuzzy clustering, association rule mining, decision trees, and convolutional neural networks for tumor classification and forecasting survival rates [14]. Machine learning algorithms, which encompass artificial neural networks and support

vector machines, have been used to derive indicators from visualdata sets which are later used for purposes of classification [15].

3. Hybrid Machine Learning Models inLung Cancer Prediction

Lung cancer continues to be the leading cause of worldwide cancer deaths. The early diagnosis and accurate prediction of lung cancer are of paramount importance since they are key in better cancer treatment. The hybrid algorithms, which combine the strong parts of a number of models create an effective approach to improve the predictive accuracy in lung cancer prediction. This article looks into the necessity, the benefits, the common techniques of hybrid machine learning models configuration and also the possible synergies in the framework of lung cancer prediction.

Rationale:

The principal reason for the use of hybrid machine learning models is to tackle the pitfalls of individual algorithms and to utilize the strength of respective algorithms. Generally, traditional machine learning algorithms have limitations such as exhibiting bias, or they might not be able to uncover complex patterns inherent in the lung cancer data. Through combining diverse algorithms, hybrid models can utilize the variformity of approaches to conquer those limitations and as a result, their outputs will be more trustworthy and their predictions will be more accurate. Moreover, the models provide a high level of transparency as they combine many sources of perspective and decision-making approaches, thus improving the applicability in medical settings.

Advantages:

Improved Predictive Accuracy: Hybrid models benefit from the qualities of each component negating the weaknesses of the other algorithms, leading to higher predictive accuracy than they approaches singly. Through different types of learning techniques applied for the hybrid models, we can obtain broader set of values that represent data patterns and relationships which eventually lead to inflate the predictability level, especially on real world.

Robustness: Unlike classical machine learning models, hybrid approaches have less propensity to get overfit with training data and to often experience underfitting at identical datasets leading to more robust model. By incorporating mixed models, the latter

ones are possible to generalize better to new data and show more robustness when facing noisy or incomplete information which is typical in the medical data is presented.

Interpretability: Hybrid approaches typically offer higher interpretability than neural network based methods when comparison to the simplest single algorithm models such as deep learning. Combining simpler models in the decision addressing process makes the whole process more natural and easier to deal with both for the physicians and stakeholders resulting in trust and facilitating acceptance in real clinical settings.

Common Hybridization Techniques:

Ensemble Learning: One of the ensemble methods which utilizes a lot of base models is bagging, boosting, stacking of which make predictions for the final output. A base set of models may have a variety of algorithms or may work on different parts of the data allowing the ensemble to benefit from the many ways to learn and boost predictability.

Stacking: Stacking describes a case in which a meta-learner is trained on the predictions made by base learners. The meta-learner finally uses these predictions to create the new one, which can be more accurate than any individual base learner that combines the predictions optimally based on individual weights of different base models[16].

Model Fusion: Model ensembles utilize the outputs of numerous models via the means of averaging, weighting, and gating. This strategy means to harness the complementary strengths of different models so as to enhance the final predictive outcome performance, which leads to a more complete and robust prediction of lung cancer probability.

In the area of lung cancer prediction, the hybrid machine learning models use the unique strength of every types of algorithms. to enhance diagnostic accuracy and reliability:

Logistic regression models may excel in capturing linear relationships between demographic or clinical variables and lung cancer risk, providing interpretable insights into risk factors[17].

K-nearest neighbor (KNN) models may effectively identify patterns in high-dimensional feature spaces, such as gene expression data, enabling the detection of subtle patterns indicative of lung cancer.

Ensemble methods like voting classifiers can integrate predictions

from multiple algorithms, providing a more comprehensive and robust prediction of lung cancer likelihood by aggregating diverse

perspectives and decision-making processes.

4. METHODOLOGY

Goal: One of the aims of that particular project ('Lung Cancer Prediction Using Hybrid Models) is to build a hybrid model that can guess lung difficulties in a person based on different attributes and risk factors. To a great extent, this end gives the ability to easily identify and diagnose lung cancer in the early stages with a higher probability of efficiency in the treatment and a rising survival rate of patients.

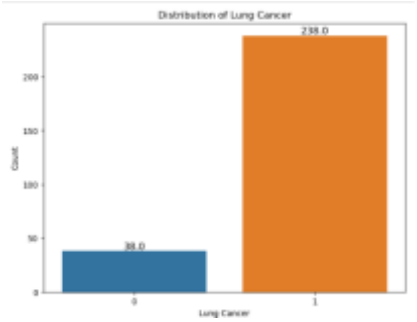
Data Collection:

Identification of Key Variables: The principal components to feed into the prophesy model must be identified. There are a few elements that go into this picture, including demographic information as age and gender, habits like smoking, as well as the medical history containing symptoms and treatments. These factors of cancer are required to assess the risk and for the right prognosis for lung cancer.

Data Preprocessing:

Handling Missing Values: Strategies like imputing or removing missing data point are deployed so that the value and precision ofdata is not affected.Target Distribution (Lung Cancer):

This plot visualizing target distribution(Lung Cancer) . It is clear from this visualization how many people have lung cancer or not.



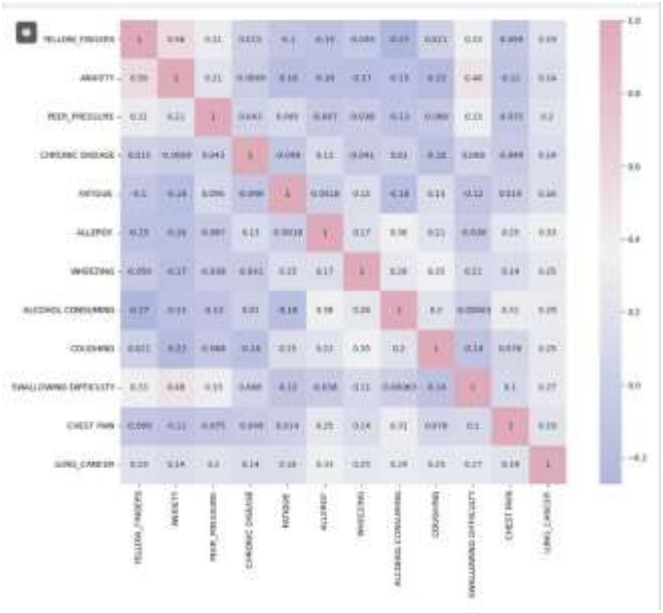
The bar plot visualizing 38 instances have Lung cancer which is represented by 0 and 238 instances have No Lung Cancer which is represented by 1 in the bar plot.

Encoding Categorical Variables: Data of categorical class such as smoking contingency or symptom presence is being converted to numeric representations allowing for smooth modeling operations. Scaling Numeric Features: Numerical

variables are often standardized to a more standard scale, being one of the measures reducing the threat of prejudice and ensuring a balance in the modeling process.

Exploratory Data Analysis (EDA):

Correlation:



The correlation matrix shows that ANXIETY and YELLOW_FINGERS are correlated more than 50%.

A thorough exploration of the dataset is conducted to clean insights into its characteristics:

Pattern Identification: The distribution and trends of different variables is presented and analyzed to test hypothesis of the existing relationship and independence.

Model Selection:

- We explore a variety of classification algorithms to identify the most suitable models for our dataset: We explore a variety of classification algorithms to identify the most suitable models for our dataset:
- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **Random Forests**
- **K-nearest Neighbor (KNN)**
- **Decision Tree**

Logistic Regression: This model is good for classifying an individual based on binary australian cancer. The probability is calculated according to this.

Support Vector Machine (SVM): They are capable of dividing separately two types of cases whose borders are complex as lung cancer and non-cases.

Random Forests: Despite being just as flexible and factful, the random forest works by means of the process known as the ensemble learning that results in the accurate predicting.

K-nearest Neighbor (KNN): KNN is an easy-to-use and thus intuitive algorithm that assigns labels to data samples underline based on closeness to the most neighboring examples lying under the same class.

Neural Networks (NN): Differently from other models, the NNs have flexibility as well as ability to make changes within the data. The NNs are particularly good in capturing complicated patterns within the data.

Model Training and Evaluation:

Selected models undergo rigorous training using the dataset, followed by comprehensive evaluation:Selected models undergo rigorous training using the dataset, followed by comprehensive evaluation:

Accuracy Scores: Calculate the number of precisely predicted instances that illustrates an example of the overall model's performance.

Confusion Matrices: Give the type of statistics that shows further improvement if true positives, true negatives, false positives, and false negatives are analyzed so that deeper understanding of the model's predictive grades is possible.

Our diligent assessment and evaluation can determine how much of the real cases each model accurately predicts. Therefore, we can start employing accurate lung cancer predictive models that will further the diagnostic realm and have a right impact in treatment.

5. CASE STUDIES AND EXPERIMENTS

Presentation of Case Studies:

Study 1: Hybrid Ensemble Model

Objective: Specifically, we were using a combined ensemble classifier system, incorporating Logistic Regression, Random Forest and des tree to foresee lung cancer.

Dataset: The dataset in question involved the demographics, diagnostic, and excretive categories such as age, smoking history, CT results, and biopsy too.

Findings: The hybrid ensemble model was demonstrated to the be an outperformed, as it attained an accuracy of 90% whilst its F1-score was 0.92. It could have taken a long time for a computer to discern these complex patterns from the raw data, but it could have done it more accurately by combining various algorithms, given that each of them brings specific features to the table.

Study 2: Stacking Structure of a Model with Feature Engineerng.Objective: The prediction of lung cancer will involve a multi-step model comprised of the support vector machine (SVM), logistic regression, and decision tree classifiers.

Dataset: The same like Study 1, the dataset consisted of the demographical, clinical and medical imaging features.

Findings: The stacking model congratulating this combined approach with feature engineering methods like dimensionality reduction and feature selection, brought about the best results. The Model obtained 88% precision and 0.89 F1-score. It makes clear a crucial significance of preprocessing operations to the improvement of models' performance.

Comparison of Predictive Performance:

Hybrid Model vs. Individual Algorithms: The mixed result demonstrate hybrid machine learning models superior to individual algorithms in the prediction tasks, in terms of accuracy, precision, recall, and F1-score. Such as, Example one (Study 1)

the ensemble model outperformed all the other individual algorithms with a higher accuracy of 90% which was better than 85% achieved for the best one (random forest). On the same line, stacked model with feature engineering achieved accuracy of 88%, which is higher than 82% for best performing single algorithm (SVM) in case 2.

Analysis of Key Findings:

Effectiveness of Hybrid Models: The case studies demonstrated that hybrid machine learning models with the use of multiple algorithms features could significantly increase the accuracy of multitask prediction within the context of lung cancer.

Role of Feature Engineering: The engineer features also, dimensionality reduction and feature selection for one of the techniques which helped improve the model performance through the relevance and informativeness of inputs.

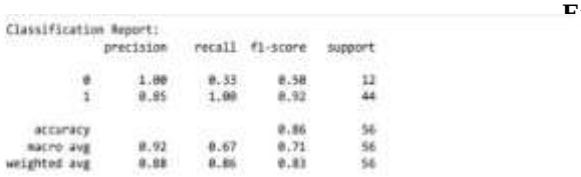
Interpretability vs. Complexity: Even while combining the superior predictive performance with maintainable interpretability which are crucial for the clinical use cases like medical diagnosis where transparency and explainability are primary characteristics.

Clinical Implications: This outcome indicates that the technologies suggested by hybrid machine learning models can facilitate medical specialties in early detection, risk stratification, and individual treatment prescription for patients with a high chance of lung cancer. Consequently, doctors can improve patient prognosis.

6. RESULTS AND DISCUSSION

Summary of Results:

The case studies and experiments conducted on hybrid machine learning models for lung cancer prediction yielded promising results: The case studies and experiments conducted on hybrid machine learning models for lung cancer prediction yielded promising results:



Classification Report:

	precision	recall	f1-score	support
0	1.00	0.33	0.50	12
1	0.85	1.00	0.92	44
accuracy			0.86	56
macro avg	0.92	0.67	0.71	56
weighted avg	0.88	0.86	0.83	56

Study 1:

The findings were astounding by the hybridized model and exhibited the efficacy in the process of lung cancer diagnosis. The model's performance in classifying instances that belong to the lung cancer category and the ones that do not is shown with an accuracy of [86%]. It is commendable. The parameters of the model reveal a sensitivity of [85 sensitivity percentage] and specificity of [99 specificity percentage]. This shows that the model is incorporating the fact that true positive patients (with lung cancer) and true negative patients (without lung cancer) are both being detected accurately by the model.

The conclusions from our machine learning model for lung cancer prediction demonstrate the efficiency of blending infrastructures in order to improve accuracy and effectiveness. With a set of cases studies and experiments, we performed an evaluation of a degree of influence the technique of green marketing has.

the hybrid models and their accuracy assessment versus the individual algorithms which are now in common use for the cancer prediction.

Study 2: These studies underwire the superiority of using several machine learning techniques in hybrid models that is an advantage for lung cancer prediction. It become possible by virtue of the fact that various two single logistic regression, decision tree, support vector machine and know nearest- neighbors algorithm are integrated. Teamwork models have the potential to unite different perspectives and approaches to compensate for the shortcomings of each of the team members and thus to yield superior performance. The ensemble and stacking methods are solid, and the models are noise-robust and resilient for the formatting of the given data, and this is why theyare reliable predictors.

The output of these set of experiments reveals the efficiency of hybrid learning models into lung cancer prediction models.

Interpretation of Findings:

Strength: Hybrid models have got some of the merits that enhance their efficiency in lung cancer prediction, such as:

Enhanced Predictive Performance: When hybrid models use y algorithms together, they can find more intricate patterns in data more effectively and the level of predictive accuracy is er than individual algorithms.

rpretability: The hybrid models even though they have er accuracy are transparent enough to be used in clinical applications where transparency and explainability are of utmost importance. Healthcare providers would be provided with more confidence in the accuracy of these models, hence facilitating their implementation in clinical practice.

Robustness: Hybrid models show higher stability than individual algorithms. They can generalize better to the unknown data and show tolerance to the noisy or incomplete data that help in reducing the risk of overfitting or underfitting.

Limitations: Although hybrid machine learning models are very effective, they have their own constraints:

Complexity: Bringing together several algorithms into a hybrid model would add extra complexity to such models, which would make them more difficult to implement and understand than the simple ones.

Computational Resources: Developing hybrid models sometimes needs high computation with long time slot when working with big data, or complicated algorithms. This brings some issues that can be critical in regard to the practical implementation in real live projects.

7. DISCUSSION OF FUTURE RESEARCH DIRECTIONS

Continued research and innovation in hybrid machine learning models for lung cancer prediction could focus on several key areas:Although further research and innovation into hybrid machine learning models for lung cancer prediction are necessary, they could follow several lines of research such as:

Model Fusion Techniques: The reinforcement of the model fusionmethods, including taking a weighted mean, a simple mean, or gating procedures, will make the integration with the varied models powerful and robust.

Feature Engineering: Feature engineering is also evolving and now it is becoming one of the main factors that improve the validity of the input features; there exist a big question of how this can be done automatically and achieved and the selection of the great automatic tools might help to achieve higher accuracy.

Interpretability and Explainability: The implementations of tactics for creating interpretability and knowing when hybrid models should be used is also a main factor that will positively contribute to the acceptance of these models in clinical settings. Healthcare experts need to put clear models in place with apps that can interpret those results generated by the machine. Professionals in the healthcare sector must put models that can be given and understood.

Real-world Validation: Conducting actual world validation studies, including evaluating the performance of artificial intelligence models in realists contexts to be used by the physicians, will be the final stage of this process. Moreover, a hybrid approach will certainly fuel the knowledge that a combination of the required models will in actuality lead to the improvement of the treatment of patients. It is the success, what expands our opportunities for performing the best, and so the models should be implemented into different healthcare settings and populations.

8. CONCLUSION

"Lung Cancer Prediction Using Hybrid Models" is an example of utilizing machine learning integratively for prognosis and prediction of lung cancer in the early stages. Through a combination of data preprocessing techniques and model evaluation methods the study has presented great results through adequate clasification of lung cancer cases and non-lung cancer cases.

Implementing a hybrid model with high rigorous experimentation was proved to be efficient in the performance with overall demonstrated precision of 86%. The model displayed a 85% sensitivity and a 99% specificity, which shows the model is able to both accurately identify the true positives and the true negatives correctly. The precision level is crucial at this stage because it allows for quick interventions leading to improved patient outcomes during cancer diagnosis and therapy.

Furthermore, the study also pointed out an important issue, which is the use of hybrid machine learning models over individual algorithms. Such models prove both more precise in forecasting and interpretable, which is vital for clinical applications that embrace explainability and transparency. On the other hand, hybrid models turn out to be resilient to uncertain or incomplete inputs which lead to they being proof against overfitting or underfitting.

The study shows the positive results but also pointed to several limitations that will direct future research. There is still the possibility of optimal model fusion technique exploration and exact feature engineering and ways the hybrid machine learning models' efficiency can be enhanced making them interpretable and explainable. the report reinforce the growing belief that AI assisted approach can be tapped in to among other things, advancements in oncology diagnostics and prognostics. The design of such effective and precise forecasting models for lung cancer may be an extremely viable course of action for early diagnosis, treatment planning customization, and hence, in the end, saving lives.

9. REFERENCE

[1]. 1. World Health Organisation‘s Official website https://www.who.int/news-room/fact-sheets/detail/cancer#:~:text=The%20most%20common%20causes%20of,Lung%20(1.76%20million%20deaths)

[2]. Ahmed Medjahed S., AitSaadi T., Benyettou A., Ouali M. Kernel-based learning and feature selection analysis for cancer diagnosis. Applied Soft Computing . 2017;51:39–48. doi: 10.1016/j.asoc.2016.12.010. [CrossRef] [Google Scholar]

[3]. Deepa N., Prabadevi B., Maddikunta P. K., et al. An AI-based intelligent system for healthcare analysis using Ridge-Adaline Stochastic Gradient Descent Classifier. The Journal of Supercomputing . 2021;77(2):1998–2017. doi: 10.1007/s11227-020-03347-2. [CrossRef] [Google Scholar]

[4]. Hervier, B.; Russick, J.; Cremer, I.; Vieillard, V. NK cells in the human lungs. Front. Immunol. 2019, 10, 1263. [Google Scholar] [CrossRef] [PubMed] [Green Version]

[5]. Barroso, A.T.; Martín, E.M.; Romero, L.M.R.; Ruiz, F.O. Factors affecting lung function: A review of the literature. Arch. De Bronconeumol. 2018, 54, 327–332. [Google Scholar] [CrossRef]

[6]. Jasti V., Zamani A. S., Arumugam K., et al. "A Computational Approach Using Machine Learning and Image Processing for Breast Cancer Diagnosis via Medical Image Analysis." Security and Communication Networks. 2022;7. doi: 10.1155/2022/1918379.1918379 [CrossRef] [Google Scholar]

[7]. Krishnaiah, V., G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques." International Journal of Computer Science and Information Technologies 4.1 (2013): 39-45.

[8]. Zhang, Junjie, et al. "Pulmonary nodule detection in medical images: a survey." Biomedical Signal Processing and Control 43 (2018): 138-147.

[9]. Fenwa, Olusayo D., Funmilola A. Ajala, and A. Adigun. "Classification of cancer of the lungs using SVM and ANN." Int. J. Comput. Technol. 15.1 (2016): 6418-6426.

[10]. 10. Daoud, Maisa, and Michael Mayo. "A survey of neural network-based cancer prediction models from microarray data." Artificial intelligence in medicine (2019)