# Improving Lung Cancer Prediction with Machine Learning Techniques

**Dr.V.R.Sonawane**

*Department of IT, MVPS's K.B.T College of Engineering, Nashik, 422013, India*

*E-mail: vijaysonawane11 @gmail.com*

**Dr.R.G.Dabhade**

*Associate professor, Department of Electronics and Telecommunication, Matoshri college of Engineering and Research center, Nashik, India.*

*Email: rgdabhade@gmail.com*

**Dr. Sumit Bhattacharjee**

*Principal, Department of CSE, Affiliated to university of Mumbai, R. V. Patel College of Science, Commerce and Arts*
*Email: bsumit021@gmail.com*

**Dr. Musmade Bhausaheb Bhanudas**

*Professor in Instrumentation and Control Engineering, D Y Patil College of Engineering Akurdi Pune-411044*
*Email: bbmusmade@dypcoeakurdi.ac.in*

*Abstract. Lung cancer is one of the leading causes of cancer-related fatalities worldwide, and the disease is common in India. Its symptoms usually show only in advanced stages, making it difficult to identify, resulting in a high mortality rate compared to other cancers. As a result, there is a need for early lung cancer prediction in order to diagnose it, which can lead to a better probability of successful treatment. Using image processing tools, histopathology images from a lung scan can be utilised to classify lung cancer. The technology extracts information from lung scans and uses them to make predictions. This article provides key step in improving classification is feature selection, which tends to provide crucial features that aid in accurately and efficiently distinguishing between various classes. As a result, optimum feature subsets can considerably increase classifier performance. The improved stochastic diffusion search (SDS) technique is used to present a unique wrapper-based feature selection algorithm. In order to discover appropriate feature subsets, the SDS will benefit from direct agent contact. For classification, neural networks, Nave Bayes, and decision trees were utilised. The experimentation shows that the suggested strategy outperforms existing methods such as lowest redundancy, extreme relevance and correlation-based feature selection.*

***Keywords: Machine Learning, Feature Selection, cancer, SDS***

## 1. Introduction

Lung cancer identified as leading cause of cancer death, responsible for almost 25% of all deaths. It is extremely difficult to detect it early because most symptoms appear during later stage. Comparatively lung cancer is more dangerous than breast cancer, colon, or prostate cancers. Various techniques such as Sputum Cytology, Chest Radiograph (x-ray), Magnetic Resonance Imaging (MRI), and Computed Tomography (CT) are used to discover lung cancers which are more expensive.

To diagnose and treated successfully there is need of early cancer detection. To automate the medical field results in an interest to use computation such as Machine learning (ML) and artificial intelligence (AI). ML comprises analysis and interpretation with the help of algorithms to provide insights and correct predictions. Actually ML concept is older one but now days many applications showing lot of potential for use in cancer treatment [1].

In this paper novel image processing technique "radiomics" and machine learning methods are discussed. Radiomics performs data mining on image datasets predict clinical and phenotype gene data. It proceeds with training (feature selection) phase followed by testing (application phase). In first phase all features are extracted from large collection of training data where target object like tumours is defined as computer program which automatically extract quantitative features [2].

Finally, a feature selection is used to select a subset that may be smaller, having features that effectively capture the image characters that are normally found within their phenomena in their biology. In a testing stage, radiomics are applied to a patient's image, and this will be a procedure, similar to a training phase, with the

selected features recognised by the Algorithm. However, in order to translate radiomics and their values into classifications, training and testing steps of an Algorithm must be specified.

Four important steps are involved while detecting lung cancer using CT images. Several approaches are used at each step, each with different accuracies in detecting lung cancer. The first stage is to pre-process the lung CT picture to remove the noises. the second stage is to segment the image to obtain the Region of Interest (ROI). The third stage is to use feature extraction to extract features like entropy, energy, variance, and lastly classification algorithms are applied to the features of the lung tissue retrieved from the CT image of the lungs.

## 2. Literature Survey

The grey level co-occurrence matrix and Gabor filter feature extraction were used by author [4]. Feature selection, which leads to the presentation of crucial features that aid in precisely and efficiently distinguishing between several classes, was an extremely critical stage in enhancing the classification. As a result, selecting the best feature subsets can help the classifier perform better. With the use of a modified Stochastic Diffusion Search (SDS) Algorithm, a new wrapper-based feature selection algorithm is proposed in this article. For identifying optimal feature sub-sets, the SDS benefits from direct agent contact. The classification algorithms were Neural Network, Nave Bayes, and Decision Tree.

For decreasing the difficulties in the feature set, author [5] introduced an effective and optimised neural computing and soft computing technique. The ELVIRA Biomedical Data Set Repository was used to gather lung bio-medical data first. Bin smoothing normalisation was used to eliminate noise from the data. The dimensionality and complexity of the characteristics were subsequently minimised by using minimum repetition and wolf heuristic features. AdaBoost optimal ensemble learning generalised neural networks were used to analyse selected lung properties. They successfully examined biomedical lung data and classified normal and pathological aspects with high accuracy. The efficiency of the system was then evaluated using a MATLAB setup for error rate, precision, recall, G-mean, F-measure, and prediction rate.

Author [6] used machine learning algorithms to predict lung cancer histology and metastasis using radiomics-based CT characteristics. The procedure required retrospectively analysing local imaging datasets of proven primary malignant lung tumours for testing and validation. CT scans were segmented semi-automatically using the same way. The clinical and computer parameters of intensity, texture, shape, and histogram were used to distinguish tumours. For the analysis, three machine learning classifiers each used up to 100 different features. While avoiding whole-body image scanning, the radiomics characterisation method showed great promise for use as a computation model in identifying lung cancer histological subtypes and metastatic forecasts for therapy decision support.

Fisher and ReliefF feature selection algorithms, as well as BP Neural Networks, were used to offer a technique for predicting lung cancer risk. To forecast the risk of lung cancer, some risk indicators were chosen. First, identify the risk variables using a mix of the two feature selection algorithms, and then use the neural network to produce the predictive value. The provided algorithm is LCRP, which is based on a framework for minimising the number of risk variables obtained in practice. The technique proposed by author [7] is suitable for both health monitoring and self-testing. The results of the experiments showed that appropriate accuracy may be achieved with smaller risk variables.

Author [8] investigated the ability of multiple machine learning classifiers to reliably identify the status of lung cancer nodules while simultaneously accounting for the associated false positive rate. A total of 416 quantitative imaging biomarkers were extracted from CT images of 200 patients' lung nodules, which had been classified as carcinogenic or benign. Biomarkers for imaging were extracted from nodule and parenchymal tissues. For classifying the binary findings of malignant or benign, linear, non-linear, and ensemble predictive classification models, as well as other feature selection strategies, were used. The top classifiers were elastic net and SVM, as well as a linear combination or correlation features selection technique, whereas the worst were RF and bagged trees. The false-positive rate for the former was roughly 30%, which was lower than the NLST's estimate. Radiomic biomarkers combined with machine learning approaches show promise as a tool for classifying cancers with low false-positive rates.

Author [9] developed Cascaded Wx, a new prognosis-related feature selection approach (CWx). Through the training of neural networks with three distinct high and low-risk groups in a cascaded manner, the CWx sorts features based on the survival of a given cohort. The concordance index showed that the top 100 genes discovered by CWx performed better or on par with those chosen by other approaches in terms of prognostic capacity. Furthermore, these top 100 genes were discovered to be linked to the Wx signalling pathway, providing

biologically meaningful evidence for their utility in predicting the prognosis of LUAD patients. More research with different cancer kinds validated the method's effectiveness. Overall, this methodology holds a lot of promise for finding prognosis-related biomarkers in the future.

Author [10] proposed a method for selecting ensemble characteristics based on t-tests and genetic algorithms. After post- and pre-processing the data with t-tests, Nested GA is used to find the best subset of features by combining data from two different datasets. Nested GA is made up of two nested GAs that run on two different datasets. The outer GA analyses microarray gene expression data, whereas the inner GA analyses DNA methylation data. With five-fold cross-validation, nested GA is performed on a colon cancer dataset. The Incremental Feature Selection (IFS) approach is used to find the least optimal gene subset once Nested GA has been applied. A separate dataset was used to validate the gene subset, resulting in classification accuracy of 99.9%. The biological significance of the resulting optimum genes is next examined using Enrichment Analysis. Furthermore, the results of Nested GA have been compared to the results of a number of other feature selection algorithms that deal with Gene Expression or DNA Methylation datasets.

Author [11] proposed a new set of quantitative features for assessing tumour responses early during chemo-radiotherapy by capturing intensity variations in PET/CT scans over time and space. The effectiveness of novel traits combined with machine learning in predicting outcomes is tested here. The proposed method works by dividing the tumour volume into successive zones based on the distance from the tumour border. Mean intensity changes in each zone of CT and PET scans are determined individually and used as image characteristics for tumour response evaluation. Tumors are described by simultaneously accounting for both temporal and geographical changes. The unique features were assessed on 30 persons with NSCLC who had received sequential or concurrent chemo-radiotherapy using linear SVM. Two PET/CT images taken before and during the first three weeks of treatment were used to predict two-year overall survival. The recommended longitudinal pattern characteristics were compared to the preceding radiomics feature and radio-biological parameters in terms of prediction ability. The PET/CT images were used to create a novel set of quantitative image attributes focusing on core tumour physiology.

Based on 3-D non-enhanced CT and CT-enhanced (CTE) characteristics [12] discovered the best machine learning technique for pre-operative differentiation of Sacral Chordoma (SC) and Sacral Giant Cell Tumor (SGCT). A total of 95 persons were chosen, divided into two groups: training and validation. There were three feature selection and classification procedures used. A comparison of their ability to distinguish between SC and SGCT was done. AUC and ACC analyses were used to evaluate the performance of the radiomics model. The CTE features were superior to the CT traits. The highest performers were the LASSO and GLM classifiers, which could lead to improved sacral tumour identification and classification.

Author [13] focused on two PCP goals: binary classification for predicting whether a person would experience a postoperative problem and 3-class multi-label classification for predicting the postoperative complication the person will experience. Furthermore, retrieving crucial data from digital medical records is an important requirement of PCP. For lung cancer PCP, a new multi-layer perceptron model called medical MLP is proposed, as well as the gradient-weighted class activation mapping Algorithm. The proposed medical MLP, which has one locally connected layer and completely connected layers with a shortcut link, extracts essential variables while also performing PCP tasks. The testing revealed that medical MLP performed better than the standard MLP. The results of the experiment showed that medical MLP outperformed regular MLP on both tasks and outperformed existing feature selection approaches. It was established using medical MLP that the "period of indwelling drainage tube" was related to lung cancer post-operative problems.

The approaches developed by [14-15] provide an excellent tool for predicting lung tumour categorization and have a role to play specifically in the finding and classification of medical data. Several lung cancer diagnosis algorithms have been discovered that use SVM to predict both normal and abnormal lung tumours. The study focuses on determining if lung cancer is normal or abnormal using categorization methods. The appropriate information from the input dataset is extracted first in the pre-processing stage. The resulting output is delivered to the features selection after pre-processing. The features are picked with the help of the firefly Algorithm in this phase. The SVM classifier is supplied with certain features.

Author [16] used the streaming features Algorithm to perform causal discovery and causal discovery with symmetrical uncertainty. It differs from traditional learning algorithms, which obtain all compute features in advance and then select the best subset of features from them. The proposed method combines the selection of online streaming features with causal structure learning. The dynamic selection of computational features, the evaluation of feature subsets, and the implementation of causal structure learning are the primary issues. Furthermore, creating a causal structure network is a time-consuming process that can be sped up by using SVM

based on the streaming feature Algorithm. The results of the experiment show that the proposed algorithm is superior to existing algorithms in terms of performance.

## 3. Motivation

Lung cancer is one of the leading causes of cancer-related fatalities worldwide. The patient's medical history and histological classification of lung cancer have offered crucial information about tissue features and anatomical locations. The radiomic characteristics and their ability of prediction in the identification of lung cancer have been presented in numerous research. However, its quantitative magnitude in terms of data is vast, posing significant hurdles to categorization algorithms. To address this, a symbolic method to data analysis is presented, which makes use of a wide range of quantitative data.

The research also looks into several feature selection techniques for predicting lung cancer histologic subtypes using either symbolic data or radiomic characteristics. These features were extracted using a Grey Level Co-Occurrence Matrix (GLCM), the Gabor filter, and fusion, which was accomplished by concatenation after the Z score was normalised. Using the modified Stochastic Diffusion Search (SDS) Algorithm, a novel wrapper-based feature selection algorithm is proposed. In order to discover optimal feature subsets, the SDS will benefit from direct agent contact.

## 4. NP Hard Problem

Because the feature selection is NP-hard, an optimal solution cannot be guaranteed unless an extensive search inside the feature space is performed. A typical feature selection procedure consists of four steps subset formation, subset evaluation, stopping criterion, and finally, result validation shown in figure1. The production of subsets is the first step, which has two major challenges [17].

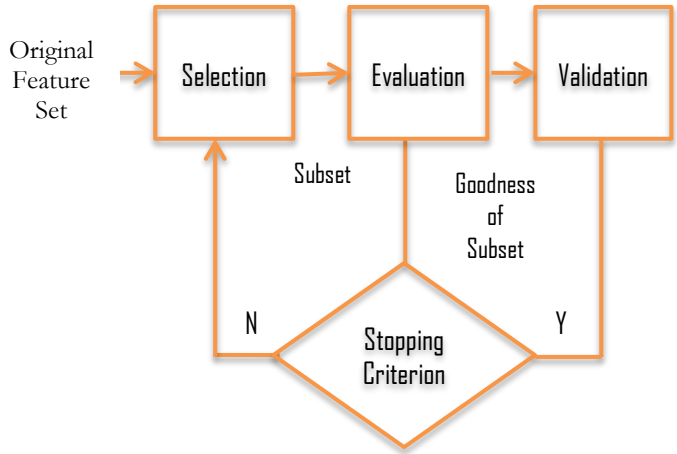


**Figure 1. Feature Selection Process with Validation**

Three important strategies are identified

1. Forward selection starts with no variables and adds one at a time for each step, each one reducing the error until there is no more to reduce.
2. Backward selection starts with the variable, which is deleted one by one for each level, reducing the one that can overcome the error the most until such time as it can no longer overcome the error. Furthermore, elimination tends to fix the errors.
3. Two ends are used to start a progressive selection. It simultaneously increases or decreases the number of characteristics.

### 4.1. Heuristic Optimization Algorithm

Two main components are very significant in the heuristics: exploration or diversification (Ability of global search) and exploitation or intensification (ability of a local search). The success of the algorithm is reliant on a perfect balance of both components. If exploration is insufficient and exploitation is excessive, a premature convergence may occur. Simultaneously, if there is too much exploration and not enough exploitation, there may

be difficulties in achieving convergence toward optimal solutions [18]. The term "Genetic Algorithm" (GA) refers to a search technique that is used to find approximate solutions to issues. Genetic Algorithms are a subset of these evolutionary algorithms that employ techniques inspired by crossover, selection, mutation, and inheritance. A gene exchange between parents is used to optimize the process. A simple GA has five steps [19].

1. Begin with a population, which is randomly produced for M chromosomes, with a population size of M and a length of chromosome x of l.
2. Calculate the fitness value for chromosome (x) of the population (x).
3. Rep this process till M offspring have been created
    a. Using the fitness function value, randomly select a new pair of chromosomes from the current population.
    b. I = 1, 2... N, generate offspring y using crossover and mutation operators.
4. Replace the existing population with the newly formed population.
5. Go to the stage 2.

## 5. Result Analysis

Matlab and Weka tool is used for the implementation of Stochastic Diffusion Search (SDS) Algorithm. The classification accuracy is computed after the testing data is loaded. The results are obtained by 10-fold cross validation. The performance of the approaches MRMR-NN, CFS-NN, SDS-Decision Tree, SDS-Nave Bayes, and SDS-NN are discussed here. Tables 1 to Table 3 demonstrate the classification accuracy, recall, precision and same is depicted in figure 2 to 4.

**Table 1. Classification Accuracy**

| Techniques | Accuracy |
|---|---|
| MRMR- NN | 87.02 |
| CFS-NN | 85.17 |
| SDS - Decision Tree | 87.40 |
| SDS- Naïve Bayes | 88.51 |
| SDS-NN | 89.60 |

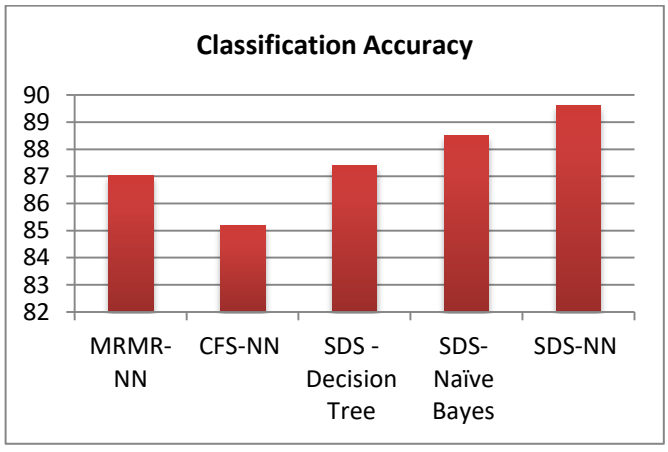


**Figure 2. Classification Accuracy for SDS-NN**

Figure 3 shows that the SDS-NN has a 0.74% higher recall for normal than the SDS-Decision Tree and the CFS-NN. SDS-Naive Bayes and MRMR-NN both have the same value. SDS-Naive Bayes and MRMR-NN both have the same value. The *SDS-NN has* a higher recall for AD than the MRMR-NN, 10.72 %for the CFS-NN, 4.74 %for the SDS-Decision Tree, and 2.82 %for the SDS-Naive Bayes. Hence because an ideal feature subset is used as input for the classifiers, the proposed SDS feature selection improves recall.

**Table 2. Classification Accuracy**

| Techniques | Normal Recall | AD Recall |
|---|---|---|
| MRMR- NN | 0.9571 | 0.7769 |
| CFS-NN | 0.95 | 0.7462 |
| SDS - Decision Tree | 0.95 | 0.7923 |
| SDS- Naïve Bayes | 0.9571 | 0.8077 |
| SDS-NN | 0.9571 | 0.8308 |

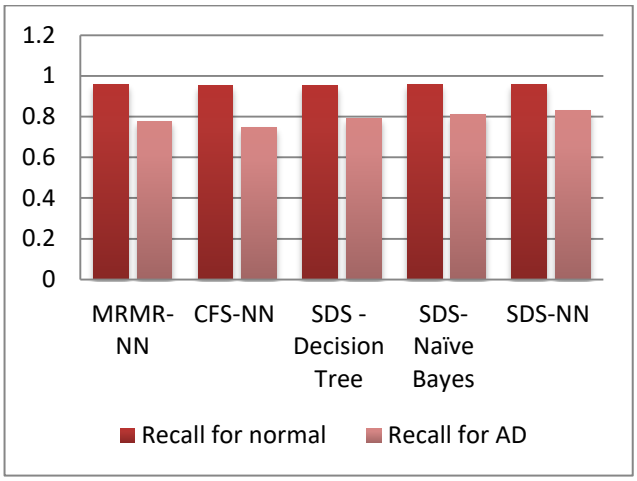


**Figure. 3. Recall for SDS-NN**

From figure 4 it is observed that SDS-NN has a greater accuracy for normal than the MRMR-NN by 4.39%, 6.96% for CFS-NN, 3.27% for SDS-Decision Tree, and 1.9% for SDS-Naive Bayes. The SDS-NN has a 0.37% greater precision for AD than the MRMR-NN, 1.56% for the CFS-NN, 1.16% for the SDS-Decision Tree, and 0.15% for the SDS-Naive Bayes.

**Table 3. Classification Accuracy**

| Techniques | Normal Recall | AD Recall |
|---|---|---|
| MRMR- NN | 0.8221 | 0.9439 |
| CFS-NN | 0.8012 | 0.9327 |
| SDS - Decision Tree | 0.8313 | 0.9364 |
| SDS- Naïve Bayes | 0.8428 | 0.9459 |
| SDS-NN | 0.859 | 0.9474 |

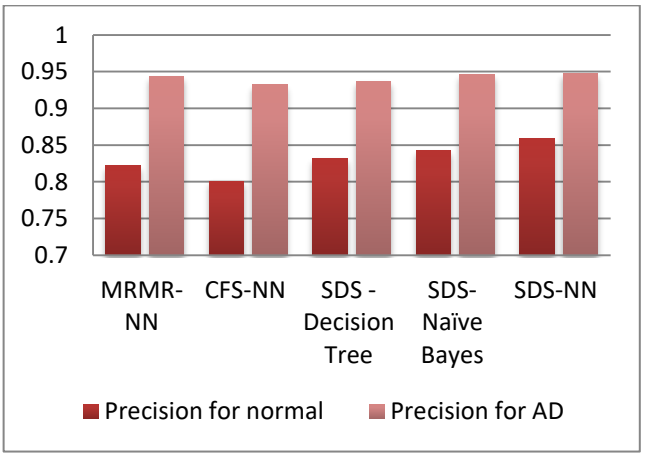


**Fig. 4. Precision for SDS-NN**

## 6. Conclusion

Lung cancer has been discovered as a high deadly disease that is widespread and the leading cause of mortality. The major goal here is to anticipate novel detection methods by using several classifiers with optimal characteristics. The predicted subsets of cancer cells that may develop in cancer were identified using a set of criteria. When certain features are omitted, however, performance improves. The SDS is used to choose appropriate subsets for classification in this strategy. The SDS has been updated for the selection of a new feature subset that is suitable in the suggested Algorithm. In terms of image classification, NNs are more efficient. It has been observed that feature selection improves image classification. More research on how to improve classifier is required. This paper focuses on feature selection, however, various pre-processing techniques such as noise removal and optimal classifiers can be examined further.

## REFERENCES

*[1]. Rabbani, M, Kanevsky, J, Kafi, K, Chandelier, F & Giles, FJ 2018, „Role of artificial intelligence in the care of patients with nonsmall cell lung cancer". European journal of clinical investigation, e12901. vol. 48, no. 4.*

*[2]. Kadir, T & Gleeson, F 2018, „Lung cancer prediction using machine learning and advanced imaging techniques". Translational lung cancer research, vol. 7, no. 3, pp. 304-312.*

*[3]. Al Rawi, MW & EL-Bakry, HM 2019, „Survey on Gene selection using meta heuristic Algorithms for Classifying Cancer Diseases".*

*[4]. Shanthi, S & Rajkumar, N 2020, „Lung Cancer Prediction Using Stochastic Diffusion Search (SDS) Based Feature Selection and Machine Learning Methods". Neural Processing Letters, pp. 1-14.*

*[5]. Shakeel, PM, Tolba, A, Al-Makhadmeh, Z & Jaber, MM 2020, „Automatic detection of lung cancer from biomedical data set using discrete AdaBoost optimized ensemble learning generalized neural networks". Neural Computing and Applications, vol. 32, no. 3, pp. 777-790.*

*[6]. Thawani, R, McLane, M, Beig, N, Ghose, S, Prasanna, P, Velcheti, V & Madabhushi, A 2018, „Radiomics and radiogenomics in lung cancer: a review for the clinician". Lung Cancer, vol. 115, pp. 34-41.*

*[7]. Xie, NN, Hu, L & Li, TH 2014, „Lung cancer risk prediction method based on feature selection and artificial neural network". Asian Pac J Cancer Prev, vol. 15, no. 23, pp. 10539-10542.*

*[8]. Delzell, DAP, Peter, T, Smith, M, Magnuson, S & Smith, BJ 2019, „Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data". Frontiers in Oncology, vol. 9, pp. 1393.*

*[9]. Shin, B, Park, S, Hong, JH, An, HJ, Chun, SH, Kang, K & Kang, K 2019, „Cascaded Wx: a novel prognosis-related feature selection framework in human lung adenocarcinoma transcriptomes". Frontiers in genetics, vol. 10, no. 662.*

*[10]. Sayed, S, Nassef, M, Badr, A & Farag, I 2019, „A nested genetic Algorithm for feature selection in high-dimensional cancer microarray datasets". Expert Systems with Applications, vol. 121, pp. 233-243.*

*[11]. Buizza, G, Toma-Dasu, I, Lazzeroni, M, Paganelli, C, Riboldi, M, Chang, Y & Wang, C 2018, „Early tumor response prediction for lung cancer patients using novel longitudinal pattern features from sequential PET/CT image scans". PhysicaMedica, vol. 54, pp. 21-29.*

*[12]. Yin, P, Mao, N, Zhao, C, Wu, J, Sun, C, Chen, L & Hong, N 2019, „Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features". European radiology, vol. 29, no. 4, pp. 1841-1847.*

*[13]. He, T, Guo, J, Chen, N, Xu, X, Wang, Z, Fu, K & Yi, Z 2019, „MediMLP: Using Grad-CAM to Extract Principal Variables for Lung Cancer Postoperative Complication Prediction". IEEE journal of biomedical and health informatics.*

*[14]. Senthil, S & Ayshwarya, B 2018, „Lung cancer prediction using feed forward back propagation neural networks with optimal features". International Journal of Applied Engineering Research, vol. 13, no. 1, pp. 318-325.*

*[15]. Senthil, S & Ayshwarya, B 2018, „Predicting Lung Cancer Using Datamining Techniques With the AID of SVM Classifier", 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, 2018, pp. 210-216.*

*[16]. Yang, J, Li, N, Fang, S, Yu, K & Chen, Y 2019, „Semantic Features Prediction for Pulmonary Nodule Diagnosis Based on Online Streaming Feature Selection". IEEE Access, vol. 7, pp. 61121-61135.*

*[17]. Samb, ML, Camara, F, Ndiaye, S, Slimani, Y & Esseghir, MA 2012, „A novel RFE-SVM-based feature selection approach for classification". International Journal of Advanced Science and Technology, vol. 43, pp. 27-36.*

*[18]. Samb, ML, Camara, F, Ndiaye, S, Slimani, Y & Esseghir, MA 2012, „A novel RFE-SVM-based feature selection approach for classification". International Journal of Advanced Science and Technology, vol. 43, pp. 27-36.*

*[19]. Chandrashekar, G & Sahin, F 2014, „A survey on feature selection methods". Computers & Electrical Engineering, vol. 40, no. 1, pp. 16-28.*

*[20]. Ahn, HK, Lee, H, Kim, SG & Hyun, SH 2019, „Pre-treatment 18F-FDG PET-based radiomics predict survival in resected non-small cell lung cancer". Clinical radiology, vol. 74, no. 6, pp. 467-473.*

*[21]. Alam, J, Alam, S & Hossan, A 2018, „Multi-stage lung cancer detection*

*[22]. and prediction using multi-class svmclassifie". In 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2) IEEE. pp. 1-4.*

*[23]. Al-Bahrani, R, Agrawal, A & Choudhary, A 2013, „Colon cancer survival prediction using ensemble data mining on SEER data". In 2013 IEEE international conference on Big Data IEEE. pp. 9-16.*

*[24]. Alhakbani, H & Al-Rifaie, MM 2017, „Feature selection using stochastic diffusion search". In Proceedings of the Genetic and Evolutionary Computation Conference ACM. pp. 385-392.*

*[25]. Ankita, R, Kumari, CU, Mehdi, MJ, Tejashwini, N & Pavani, T 2019, „Lung Cancer Image-Feature Extraction and Classification using GLCM and SVM Classifier". International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 11.*