SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I9P115

Volume 11 Issue 9, 163-173, September 2024
© 2024 Seventh Sense Research Group®

Original Article

Design of Multivariate Lung Cancer Dataset and Multistage Pre-processing to Augment Prediction of Complex Lung Cancer Datasets Using Data Mining Techniques

M. Amenraj[1], R. Vidya[2]

[1,2]St. Joseph's College of Arts and Science (Autonomous), Cuddalore, Tamilnadu, India.

Corresponding Author : amenrajm@gmail.com

Received: 05 July 2024 Revised: 15 August 2024 Accepted: 06 September 2024 Published: 30 September 2024

Abstract *- Progressions in genomic research have led to an increased focus on Single Nucleotide Polymorphisms (SNPs) as potential markers for various diseases, including lung cancer. This study introduces a novel approach to enhance the predictive accuracy of ensemble machine learning classifiers and design of a Multivariate Dataset for Lung Cancer for SNP-associated lung cancer through a three-stage pre-processing framework called Lung Cancer Data Pre-processing and Feature Engineering (LC-PreProFE). The framework comprises numerical analysis at the initial stage, followed by regression analysis and segmentation at the final stage objected to eliminate irrelevant features and optimize the construction of a multivariate dataset. The first stage involves rigorous numerical analysis to identify and quantify the significance of each SNP within the dataset. The stage eliminated 2 features with a 4% improvement in best predictions. The refined dataset undergoes regression analysis to model the relationships between identified SNPs and to filter out redundant or correlated features. This stage eliminated 4 features. Finally, in the segmentation process, 7 irrelevant features were eliminated. After completion of three stages, it was found that the accuracy has improved after irrelevant feature removal and the Region of Curve value reduced to show augmentation in the overall preprocessing stage.*

Keywords *- Single Nucleotide Polymorphisms (SNPs), Multivariate lung cancer dataset, Lung Cancer Data Pre-processing and Feature Engineering (LC-PreProFE), Ensemble machine learning models, Irrelevant feature engineering.*

1. Introduction

Lung cancer is now one of the most popular and mortal forms of cancer. Because lung cancer is very serious, many countries are now telling their citizens who are at risk to get tested and treated early. Lung cancer was more severe in poorer areas or countries, where people were more likely to get it because the local healthcare and medical resources were not enough. Over the last couple of years, significant advancements have been made in enhancing this condition by utilizing the existing knowledge regarding lung cancer in developed nations. However, the way the information was gathered was not well organized. The information gathered was different and could not be easily used. Artificial Intelligence (AI), huge information, cloud computing, and the Internet of Things (IoT) are speeding up advancements in the medical industry, the transformation being referred to as “Medical Industry 4”. Lung cancer can now be detected earlier by using a very smart method. Gu, C. et al. (2022) [1] involved the appeal of artificial intelligence and cloud platform techniques in the medical sector. Our team has developed an intelligent system proficient in locating lung cancer. It combines data, compares cases from the past, looks for similar cases, and can find information. In this system, doctors can use different types of data and put it together to have much useful information when treating the patient. A computer system used a smart model that learned from a cloud to help with finding similar images. In the end, we used some public datasets to teach and try out this system. The outcomes show that it performed better than some standard methods.

Later on, they analyzed the detection of analogous cases using cosine similarity. In all cases, the similarities measured were above zero, which combines different types of information assisting doctors and patients with lung carcinogens to have better access to diagnosis and treatment. However, it could not handle multivariate datasets.

The major objective of the research paper is to design a framework model to design a Multivariate dataset that assisted in manipulating the lung cancer dataset with SNP types through the amalgamation of pre-processing techniques and Dimensionality Reduction Techniques of data mining. The goal is to attain a solution to reduce the high dimensionality of the medical datasets and bring a solution to the problem of prediction using classifiers when different kinds and multiple sets of data are present in the input dataset for manipulations. The scope of the research work has been confined to the analysis of medical datasets that could be used to predict lung cancer with improved efficiency and performance.

2. Literature Study

In the Present Times, a lot of people's mortality has been due to lung cancer. Being able to forecast and identify early indicators of lung carcinoma has helped save many lives because it is a main cause of extinction. It is very significant to find lung cancer early to minimize the damage to the body. The use of advanced machine learning techniques was needed to catch sight of lung cancer early. Vasudha Rani. V et al. (2022) [2] desired to make a model that can estimate how likely it is for a person to develop lung cancer. The goal is to gain more information about how lung cancer develops and spreads using this activity. Lung cancer causes the most cancer deaths in both males and females. This happened because the screening programs were not very good at finding it early, and people often did not notice symptoms until the cancer had already reached advanced stages. Jaksik, R., & Śmieja, J. (2022) [3] created and tested different classifiers using mRNA and micro-RNA levels, somatic mutation positions, changes in DNA copy numbers, and DNA methylation levels.

The World Health Organization says that lung cancer is one of the extensively typical reasons why people die all over the world. Maleki, N., & Niaki, S. T. A. (2023) [4] analyzed ANN to determine what was in each picture of lung cancer images. In the following method, the pictures were prepared and divided before using CNN and ANN. Tayal, D. K. et al. (2022) [5] used a technique called "dimensionality reduction" as part of the initial processing using Grey Wolf Optimiser (GWO). Aziz, R. M. (2022) [6] helps reduce the mistakes made by the organizer and allows for quicker results by choosing the most important genes. Ramkumar, M. P. et al. (2022) [7] developed a process to improve how well anti-corona virus-Henry gas can dissolve. This is done by using an extraordinary set of statutes that find the best way and consider different things. The new method made the tests better at correctly identifying things, with a success rate of 91%.

In studying biology and computer science, the identification of different types of cancer was very important. The Microarray technology, which was widely used, helped identify different diseases. A few important genes found in medical tests could result in affordable medications that can predict how long a patient may live or detect cancer. The microarray data had many genes but not enough samples. This made high dimensionality a big problem. In this study, Venkatesan C. et al. (2022) [8] looked at the genes in the Signal-to-Noise Ratio (SNR) and studied the best genes using optimization methods to find optimal results. Hanley, C. J. et al. (2023) [9] worked on studying different types of fibroblast cells in people with non-small cell lung carcinoma using advanced techniques. Pradhan, K. et al. (2023) [10] developed a more effective method of diagnosing lung cancer by analyzing a patient's medical history. They used the SA-SLnO method to discern the adequate digit of concealed neurons in the RNN. Chassagnon, G. et al. (2023) [11] examined how AI is currently being used in thoracic oncology and what possibilities it holds for the future. The CT scans of lung cancer and tuberculosis can sometimes look the same, which can cause doctors to make a wrong diagnosis. Zhang, K. et al. (2022) [12] used a combination of profound understanding and content-based image retrieval to tell the difference between lung cancer and typical tuberculosis in CT images.

There have been a few studies that looked at cancer outcome prediction models using a type of statistical method called Bayesian hierarchical models. Sun, N. et al. (2022) [13] used a new statistical strategy to explore mRNAs and understand how they impact the outlook of patients with lung adenocarcinoma. Creating a predictive embodiment can be achieved by incorporating clinical factors and mRNAs, based on the mRNAs that have been discovered. Primakov, S. P. et al. (2022) [14] studied different aspects of a medical imaging technique, such as the thickness of the images, the size of the tumour, how difficult it was to interpret the images, and where the tumour was located.

The CAD method helped a lot in recognizing medical images of lung cancer because there are more and more patients with this disease. Pulmonary nodules are abnormalities in the lungs. They can be found early using a computer-based diagnosis tool for lung patients. The current method used to classify lung CT images has not been very good at detecting early-stage cancer and takes a long time to complete. A new method was proposed by Siddiqui, E. A., et al. (2023) [15] to improve the exactness of classifying lung CT images.

Earlier, lung disease was recognized as the worst illness, and it still is today. Catching diseases early is important to stop people from getting them. Many researchers worked together to find different ways to predict how accurate diseases can be forecasted. The machine-learned algorithm failed to predict accuracy as effectively as the deep learning technique. As a result, we suggested upgraded artificial neural system methods to better predict accuracy for lung diseases. Manoharan, H. et al. (2022) [16] used two mathematical techniques called Discrete Fourier Transform and Burg Auto-Regression to extract images from Computed Tomography (CT) scans.

Cancer used to be a major reason why many people died. It means that normal cells change into tumour cells in several different stages. Discovering carcinogens at a primary stage can greatly decrease its negative effects. Many scientific studies utilizing machine learning methods have aimed to achieve this objective. Andjelkovic, J. et al. (2023) [17] use a step-by-step method to forecast the liability of cancer, which is believed to be one of the main reasons why people die in both rich and poor countries. Categorizing cancer using the microarray dataset has helped us understand how to treat it better.

Microarray datasets had a lot of compound and excessive-dimensional genes and only not enormous number of instances. Alrefai. N & Ibrahim, O. (2022) [18] show that the suggested technique for identifying cancer using microarray data is effective, with 100% accuracy in some cases and 92% accuracy overall. Processing data becomes more difficult as the dataset becomes larger. The word "complexity" means how hard it is to find and use connections between different parts of a dataset. So, the utilization of dimension reduction helped to get rid of the complexity caused by different features. Rani R. et al. (2022) [19] examine the research on ways to improve how big data is stored and processed in different IoT applications. It reviews different techniques, discussing their benefits, characteristics, classification, and factors for assessing their effectiveness. Additionally, the article explains upcoming challenges in research and provides information on how data reduction techniques can be used in various areas, giving readers a clearer insight into its usefulness.

Prakash, P. S., & Rajkumar, N. (2022) [20] talked about using a method called dimension reduction to help classify medical data. The plan includes three parts: preparing the data, reducing the size using an adaptive artificial flora algorithm, and categorizing it. Yao, W. et al. (2022) [21] created and tested a non-invasive way to use radionics to determine how well patients with large-cell lung cancer are doing. The objective was to assess the level of ki67 declaration to determine the patient's prognosis. In this study, 120 patients with NSCLC were included.

All sufferers were sorted into two parts: one group for training (85 patients) and one group for testing (35 patients). Cancer is a really bad sickness that happens when the cells grow too fast and cannot be controlled. A Novel model was suggested by Elemam, T & Elshrkawey, M. (2022) [22] to help diagnose various types of cancer using large amounts of data. The algorithm has two parts that select features in a combined way. In the beginning, a ranker was started to put together the results of three methods that evaluate features using filters.

Scientists in the area of bioinformatics have used Artificial Intelligence (AI) methods to create computer programs that can quickly and accurately detect cancer. Gene expression analysis has proven to be very helpful in predicting what may happen with certain types of cancer. However, small sample sizes have been a problem for making powerful and useful classifiers.

In simpler terms, traditional supervised learning approaches could only work with data that was properly labelled. Because of this, a large part of microarray data sets that did not have the right information afterwards were not considered. AI-based deep-learned strategies can recognize important details from complex datasets, which shows how important they are. AI and ML were being used in biological research and healthcare, especially in studying cancer. This technology has huge benefits for improving healthcare. Some examples of what these activities can involve are finding cancer, figuring out what type it is, making treatments better, and discovering new ways to treat it with medicine. Although there is a lot of information obtainable to instruct machine learning design, it is still challenging to fully utilize artificial intelligence in cancer research and treatment. There are significant obstacles that require to be directed. The need was to use artificial intelligence to improve carcinoma diagnosis, prognosis, and therapy while also following standards. This pushed for more research in biotechnology. Gupta, S., & Kumar, Y. (2022) [23] focused on examining the utilization of artificial intelligence in recent research concerning cancer prognosis.

To create a trustworthy and quick disease diagnosis model, it was necessary to simplify the features at the start of the design process. Kar, B & Sarkar, B. K. (2022) [24] presented a technique to pick out the significant attributes from medical data to create a diagnostic model. The new method overcomes some challenges, such as limited disease specificity, information loss, and slow processing times. The method consists of two steps and combines different approaches to identify the most pertinent property for each medical dataset.

The practical calculation shows that the suggested method improved how well the datasets worked after picking out the most important information, getting rid of a lot of unnecessary and repetitive data. Lung cancer has caused many deaths in China. Liu, S., & Yao, W. (2022) [25] suggests a way to choose genes for analysis based on their KL divergence. Even though this method has some limitations, like small datasets and imbalanced data, it can help select genes that are more important for the model. Afterwards, it utilized deep neural network technology, a particular form of artificial intelligence, to generate a model. It used a technique called focal loss as a way to measure how well our model is performing.

A newly developed computer model was created by Gupta, S. et al. (2023) [26] to enable accurate identification of whether ultrasound scans of the breast are with cancer using a changed RESNET50 design that was first trained on a dataset called image net. Hambali M. A. et al. (2022) [27] suggested a new method called info gain-MBA to choose important features from large cancer datasets.

3. Methodology

A multivariate dataset in the context of lung cancer refers to a collection of data that includes information on multiple variables or attributes related to lung cancer. These variables can include various factors such as age, smoking history, genetic markers, tumour size, treatment type, survival rate, and more. Each observation or data point in the dataset represents a unique case or patient. Predicting lung cancer typically involves analyzing various factors, including patient demographics, medical history, imaging studies, genetic markers, and environmental exposures, as shown in Figure 1.

As indicated in Figure 1, Machine learning models can utilize these parameters to predict the likelihood of lung cancer or assist in its diagnosis. Some of the common parameters used in predictive models for the Multivariate lung cancer dataset are given in Table 1.

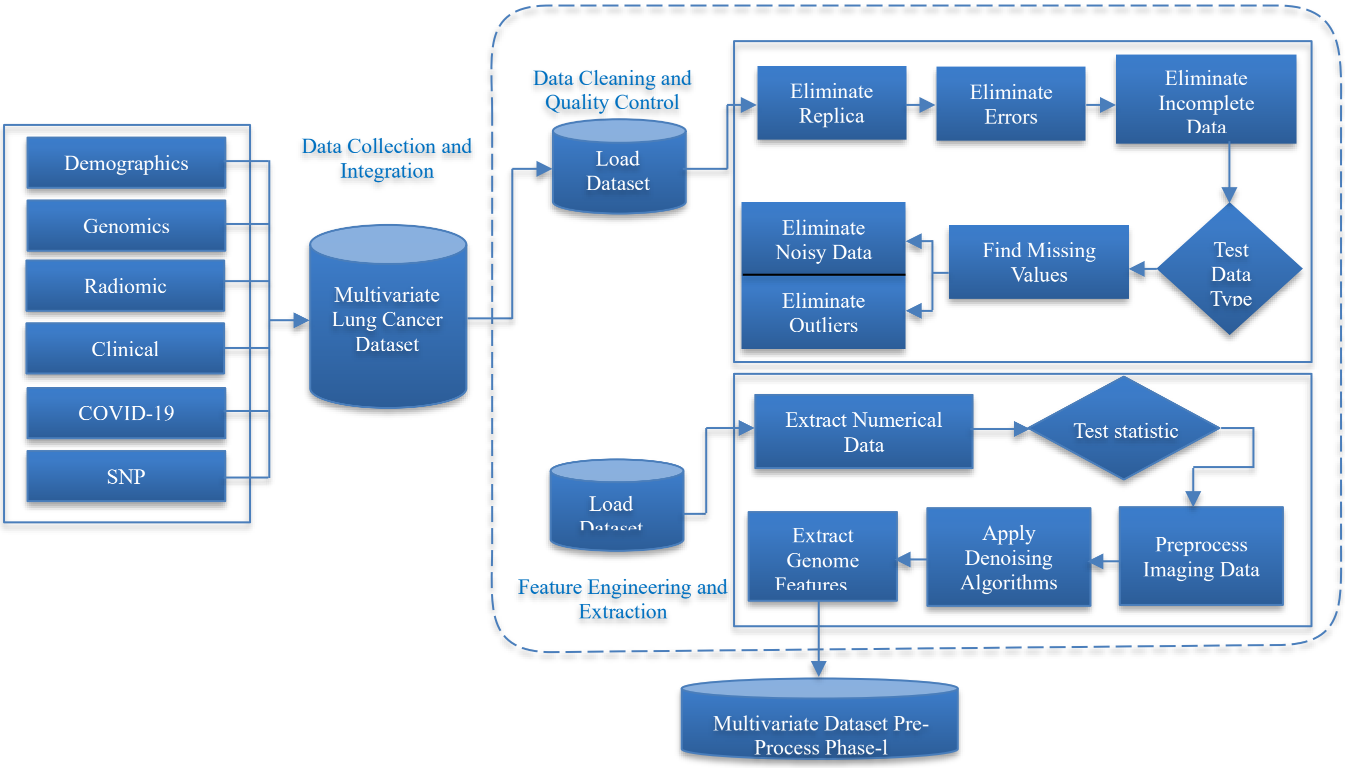


Fig. 1 Framework to prepare a multivariate dataset to predict lung cancer

Table 1. Multivariate lung cancer dataset for prediction of the disease

S. No	Feature Name	Category	Feature Type
1	Age	Demographic Information	Predictive
2	Gender	Demographic Information	Predictive
3	Smoke_History	Demographic Information	Predictive
4	Cough	Clinical Information	Predictive
5	Weight_Loss	Clinical Information	Predictive
6	History_Lung	Clinical Information	Predictive
7	History_Family	Clinical Information	Predictive
8	Lung_lesions	Xray information – Imaging	Predictive
9	Lung_nodules	Xray information – Imaging	Predictive
10	Cell_Size	CT scan information – Imaging	Predictive
11	Cell_Shape	CT scan information – Imaging	Predictive
12	Cell_Density	CT scan information – Imaging	Predictive
13	Cell_location	CT scan information – Imaging	Predictive
14	Histology_result	Pathological and Biomarker Data	Predictive
15	Genmut_result	Pathological and Biomarker Data	Predictive
16	Smoke_exposure	Environmental and Occupational Exposures	Predictive
17	Carcinogen_intake	Environmental and Occupational Exposures	Predictive
18	Radiation_exposure	Environmental and Occupational Exposures	Predictive
19	Expiratory_Volume	Pulmonary Function Tests	Predictive
20	Vital_capacity	Pulmonary Function Tests	Predictive
21	FEV1/FVC ratio	Pulmonary Function Tests	Predictive
22	Chromosome_id	Single Nucleotide Polymorphism	Predictive
23	Position_snp	Single Nucleotide Polymorphism	Predictive
24	Genotype	Single Nucleotide Polymorphism	Predictive
25	SNP_infection	Single Nucleotide Polymorphism	Predictive
26	Class_code	1 – Infected 2 – Not Infected 3 – Possible Infection	Class

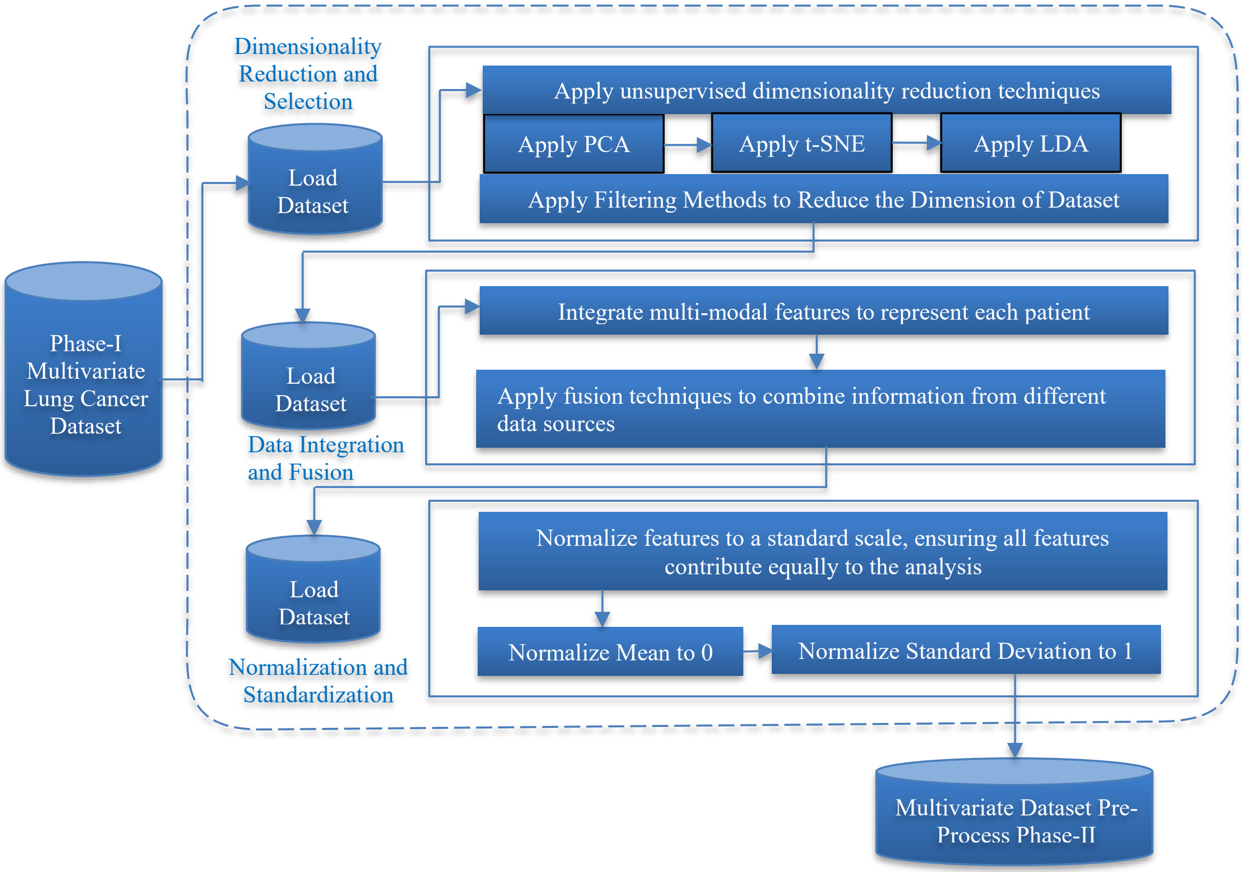

Phase-I Multivariate Lung Cancer Dataset
Dimensionality Reduction and Selection
Load Dataset
Apply unsupervised dimensionality reduction techniques
Apply PCA
Apply t-SNE
Apply LDA
Apply Filtering Methods to Reduce the Dimension of Dataset
Load Dataset
Data Integration and Fusion
Integrate multi-modal features to represent each patient
Apply fusion techniques to combine information from different data sources
Load Dataset
Normalization and Standardization
Normalize features to a standard scale, ensuring all features contribute equally to the analysis
Normalize Mean to 0
Normalize Standard Deviation to 1
Multivariate Dataset Pre-Process Phase-II

Fig. 2 Phase-I- framework: Lung Cancer Data Pre-processing and Feature Engineering (LC-PreProFE)

As shown in Table 1, the dataset comprised various Demographic Information like the Age of the individual, as lung cancer risk increases with age, Gender since Lung cancer incidence varies between males and females, Smoking history as Pack-years (a measure of smoking intensity) and smoking status (current, former, non-smoker) has serious impact on the individual. The Clinical Information comprises features like the presence and severity of symptoms such as cough, haemoptysis, weight loss, and fatigue. The medical history details like history of other respiratory conditions, lung diseases, or cancer and family history where family members with a history of lung cancer. The imaging data captured through chest X-ray features like radiological findings related to lung lesions, nodules, or other abnormalities and CT scan features like characteristics of lung nodules, size, shape, density, and location are also considered. The pathological and biomarker data, including biopsy results, Histological analysis of tissue samples and genetic markers like the presence of specific genetic mutations associated with lung cancer (e.g., EGFR, KRAS), are also used. Environmental and occupational exposures like exposure to asbestos, radon, second-hand smoke, and occupational exposures to carcinogens are also taken into consideration. The Pulmonary function tests, including lung function metrics features like Forced expiratory volume in one second (FEV1), Forced Vital Capacity (FVC), and FEV1/FVC ratio are selected. Finally, in the major work for prediction, Single Nucleotide Polymers (SNP) Biomarkers like Biomarkers associated with lung cancer risk or presence, such as CEA (Carcino Embryonic Antigen) or CYFRA 21-1, are selected as features for enhancing the design of multivariate datasets. Filter Methods is one of the pre-processing methods to evaluate each feature independently and assign a score based on statistical metrics like a chi-squared test, information gain, or correlation. Select top-ranked features based on these scores. These pre-processing methods enable a reduction in the number of features in a lung cancer dataset, making subsequent analysis and modelling more efficient while preserving relevant information for accurate predictions and insights.

4. Proposed Framework Model

Creating a novel framework for pre-processing lung cancer datasets involves devising a unique approach that optimizes data handling and feature selection to prepare the data for further analysis. The conceptual framework Lung Cancer Data Pre-processing and Feature Engineering (LC-PreProFE) is given in diagrammatic form as indicated in Figure 1.

As shown in Figure 2, various stages are incorporated in the prediction of lung cancer for multivariate datasets. They initiated data collection and integration where users collect various data sources related to lung cancer, including clinical, genomic, radiomic, and demographic data. Integrate these diverse datasets into a unified dataset for comprehensive analysis. In step 2, data cleaning and quality control are performed where the removal of duplicates, erroneous, or incomplete data is performed.

Handling missing values through appropriate imputation methods based on data type and context and conducting quality control checks to identify and handle outliers and noise are performed in this stage. Then, Feature Engineering and Extraction were performed to extract meaningful features from raw data using domain knowledge and statistical methods, leverage imaging processing techniques to extract radiomic features from medical images and utilize genetic analysis to extract relevant genetic markers and features associated with lung cancer. In the next step, Dimensionality Reduction and Selection are performed to apply unsupervised dimensionality reduction techniques like PCA, t-SNE, or LDA to reduce feature dimensions while preserving essential information and incorporate feature selection methods like recursive feature elimination, filter methods, and embedded methods to retain informative features.

In Data Integration and Fusion, the users integrate multi-modal features to create a comprehensive representation of each patient, combining clinical, genomic, and radiomic features. Later, fusion techniques were applied to combine information from different data sources effectively. The Normalization and Standardization stage normalizes features to a standard scale, ensuring all features contribute equally to the analysis and standardize variables to have a mean of 0 and a standard deviation of 1, aiding in consistent comparisons. In the Data Augmentation stage, data augmentation techniques are applied to increase the size of the dataset, especially for smaller datasets and generate augmented samples by making minor modifications to existing data, considering the specific attributes of lung cancer data. During the Validation and Evaluation stage, cross-validation is employed to assess the effectiveness of the pre-processed dataset and chosen features, and users evaluate the impact of the pre-processing steps on subsequent analysis and model performance. In the final stage, visualization and interpretability are performed to utilize visualization techniques to explore the pre-processed data, highlighting patterns and relationships and generating visualizations to aid in explaining the pre-processing steps and their effects on the data. This framework, LC-PreProFE, offers a systematic approach to pre-processing lung cancer datasets by integrating diverse data sources, optimizing feature engineering, reducing dimensionality, and ensuring data quality. It aims to enhance the accuracy and interpretability of subsequent analysis and modeling tasks related to lung cancer. The algorithm has been represented in algorithmic form in Table 2.

Table 2. Algorithm: LC-PreProFE (Lung Cancer Data Pre-processing and Feature Engineering) for pre-processing of lung cancer dataset

```
Algorithm: LC-PreProFE (Lung Cancer Data Pre-processing and Feature Engineering)

Inputs: Multivariate datasets related to lung cancer (Clinical, Genomic, Radiomic, Demographic)
Outputs: Pre-processed and integrated dataset for further analysis

Step 1: Start Data Collection and Integration
Step 2: Load the diverse datasets related to lung cancer.
Step 3: Integrate the datasets into a unified dataset.
Step 4: Perform Data Cleaning and Quality Control
Step 5: Remove duplicates, erroneous, or incomplete data.
Step 6: Handle missing values using appropriate imputation methods.
Step 7: Perform quality control checks to identify and handle outliers and noise.
Step 8: Perform Feature Engineering and Extraction
Step 9: Extract features from raw data using domain knowledge and statistical methods.
Step 10: Use imaging processing techniques to extract radiomic features from medical images.
Step 11: Utilize genetic analysis to extract relevant genetic markers and features.
Step 12: Perform Dimensionality Reduction and Selection
Step 13: Apply unsupervised dimensionality reduction techniques.
Step 14: Use feature selection methods.
Step 15: Perform Data Integration and Fusion
Step 16: Integrate features from multiple data sources for each patient.
Step 17: Perform Normalization process and Standardization of features
Step 18: Normalize features to a standard scale.
Step 19: Standardize variables to have a mean of 0 and a standard deviation of 1.
Step 20: Test for Data Augmentation after Step-17 through Step-19.
Step 21: Perform a Classifier test based on the ensemble machine learning models.
Step 22: Perform Validation and Evaluation for Step-20 through Step-21.
Step 23: Split the dataset into training and testing sets with cross validation of 10 sets.
Step 24: Train and validate models using the pre-processed dataset.
Step 25: Complete Visualization and Interpretability of Multivariate Lung Cancer dataset
Step 26: Generate visualizations to explore pre-processed data and patterns.
Step 27: Visualize the impact of pre-processing on the data.
Step 28: Return the pre-processed dataset for further analysis.
Step 29: Stop the Process

End Algorithm LC-PreProFE
```

This algorithm, represented in Table 2, provides the LC-PreProFE framework with its major implementation techniques that could assist in bringing the desired results of the research.

5. Implementation and Evaluation

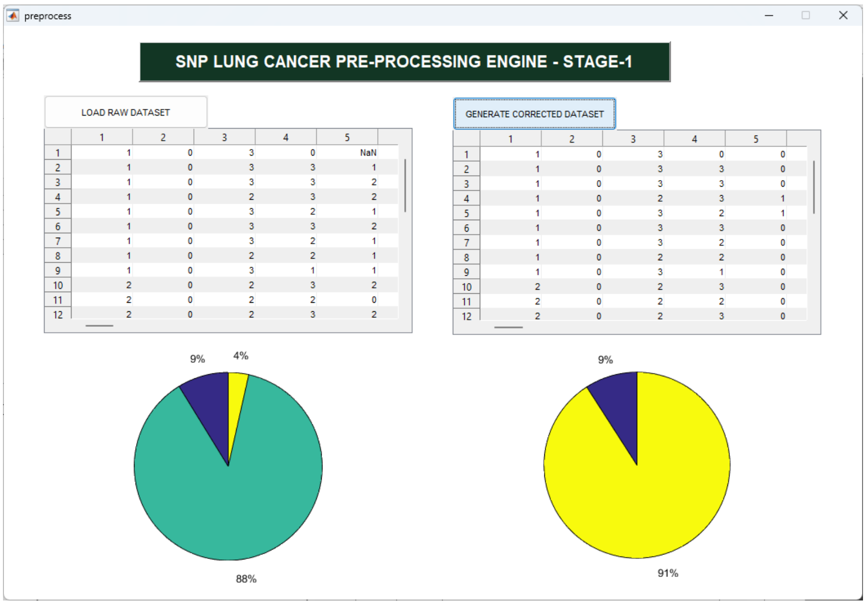
The overall research process was performed in three stages of pre-processing, as explained in distinct heads.

5.1. Stage-1 – Numerical Analysis

The initial RAW SNP Lung Cancer dataset with 57 features was tested for errors. A few data were filled with missing values and non-numeric data. The dataset comprised complete irrelevant data of around 4%, with only 88% pure data at the beginning of the evaluation process.

The numerical data and the missing values were tested in the first stage, and the values were modified. In many cases of particular columns, the values could not be changed, and such columns were identified and removed to form the stage-1 refined set, as shown in Figure 3.

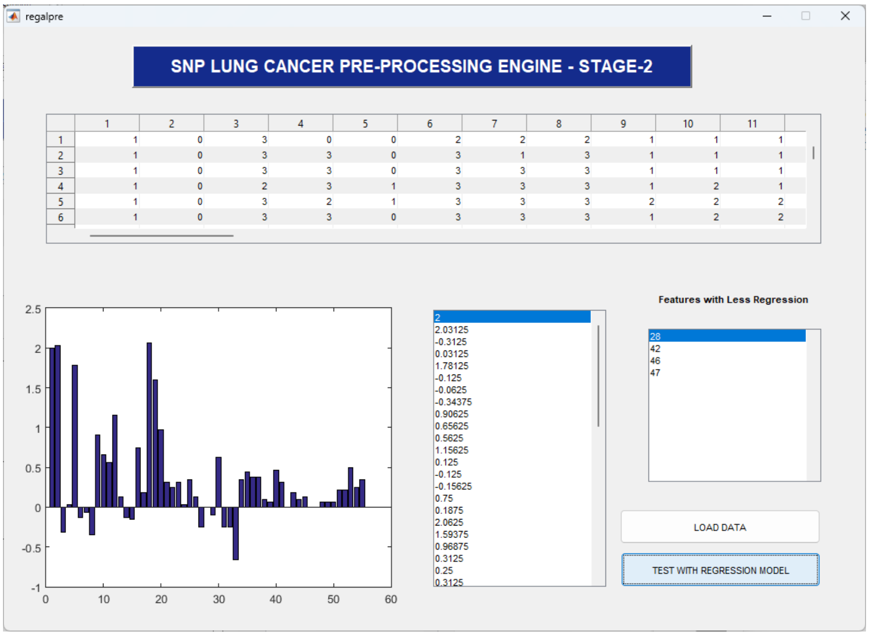
The initial stage of Figure 3 represented the graphical outcome stating that the 4% erroneous data had been removed and the percentage had improved to 91% at the end of the first stage. The completely non-numerical and missing data columns were found to be Feature2 and Feature7 and hence removed from the dataset to form 55 features at the end of this stage. After completing the process, the newly formed dataset has been stored in Excel form as a stage-1 refined set for next stage processing.


preprocess
SNP LUNG CANCER PRE-PROCESSING ENGINE - STAGE-1
LOAD RAW DATASET
GENERATE CORRECTED DATASET
9%
4%
88%
9%
91%

	1	2	3	4	5
1	1	0	3	0	NaN
2	1	0	3	3	1
3	1	0	3	3	2
4	1	0	2	3	2
5	1	0	3	2	1
6	1	0	3	3	2
7	1	0	3	2	1
8	1	0	2	2	1
9	1	0	3	1	1
10	2	0	2	3	2
11	2	0	2	2	0
12	2	0	2	3	2

	1	2	3	4	5
1	1	0	3	0	0
2	1	0	3	3	0
3	1	0	3	3	0
4	1	0	2	3	1
5	1	0	3	2	1
6	1	0	3	3	0
7	1	0	3	2	0
8	1	0	2	2	0
9	1	0	3	1	0
10	2	0	2	3	0
11	2	0	2	2	0
12	2	0	2	3	0

Fig. 3 The stage-I pre-processing of data to remove the non-numeric and missing data in SNP lung cancer data


regalpre
SNP LUNG CANCER PRE-PROCESSING ENGINE - STAGE-2
2.5
2
1.5
1
0.5
0
-0.5
-1
0
10
20
30
40
50
60
Features with Less Regression
LOAD DATA
TEST WITH REGRESSION MODEL

	1	2	3	4	5	6	7	8	9	10	11
1	1	0	3	0	0	2	2	2	1	1	1
2	1	0	3	3	0	3	1	3	1	1	1
3	1	0	3	3	0	3	3	3	1	1	1
4	1	0	2	3	1	3	3	3	1	2	1
5	1	0	3	2	1	3	3	3	2	2	2
6	1	0	3	3	0	3	3	3	1	2	2

2
2.03125
-0.3125
0.03125
1.78125
-0.125
-0.0625
-0.34375
0.90625
0.65625
0.5625
1.15625
0.125
-0.125
-0.15625
0.75
0.1875
2.0625
1.59375
0.96875
0.3125
0.25
0.3125

Features with Less Regression

28
42
46
47

Fig. 4 Stage-2 regression comparisons of models using the k-nearest neighbour model

5.2. Stage-2-Non-Regressive Data Removal

In the second stage, the relationship between the features was tested, and the regression was identified using the K-Nearest Neighbour Algorithm. Initially, the distance of the values between the features was calculated, as shown in Equation 1.

$$dist(X,X^{/}) = (\sum_{r=1}^{d} |xr - zr|\, p)^{1/p} \qquad (1)$$

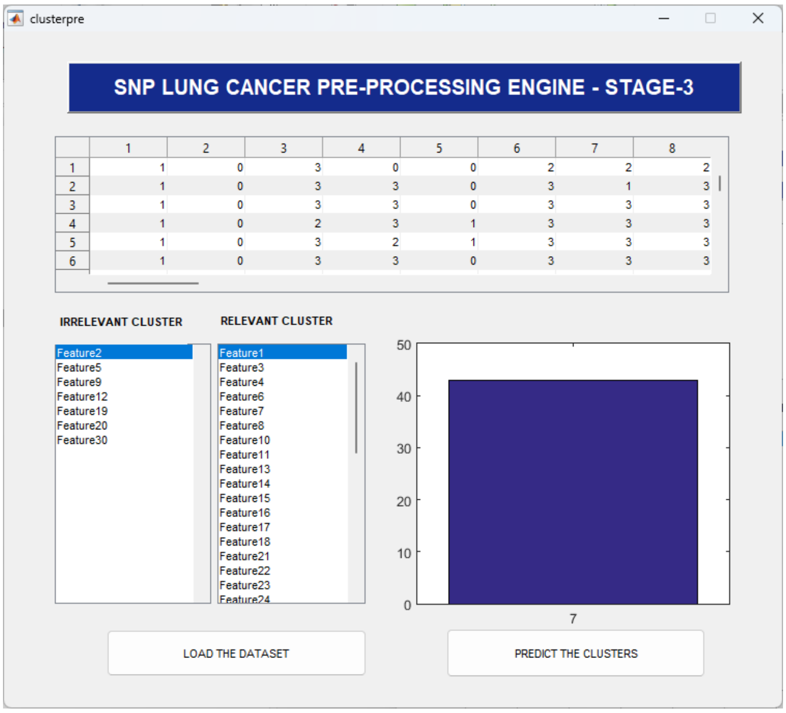
Where X and $X^{/}$ were neighbour features tested for regression and the overall summation is calculated based on the value from regression 'r' value from 1 to distance 'd'. The distance between the values has been tested and compared with the maximum distance value to indicate the threshold value calculation. The computed value is tested whether it is less than 0 or greater than 1. Based on the comparisons, the outcomes are presented as shown in Figure 4.

As shown in Figure 4, features such as Feature28, Feature42, Feature46, and Feature47 have been found non-regressive to the dataset and hence recommended for removal from the original dataset. The newly formed dataset has been formed as a regression dataset with 51 features to be created as another Excel dataset for the next stage process.

5.3. Stage-3-Segmentation Threshold Test

In the third stage, the segmentation process was performed with the remaining features based on finding the average of the feature values to form the mean. The mean value is used as a threshold value, and the features are grouped in two clusters or segments to find the relevance and irrelevance of features listed in the list box, as shown in Figure 5.

As shown in Figure 5, the 7 features, including Feature2, Feature5, Feature9, Feature12, Feature19, Feature20 and Feature30, were found to be irrelevant to the dataset and were removed from the dataset to form the refined dataset. The newly formed dataset with 44 features was created in Excel form and used as a testing set with the classifiers of ensemble machine learning models. The initial performance of the classifiers with the RAW dataset and the performance with the three-stage refined dataset is shown in Table 3.



SNP LUNG CANCER PRE-PROCESSING ENGINE - STAGE-3

	1	2	3	4	5	6	7	8
1	1	0	3	0	0	2	2	2
2	1	0	3	3	0	3	1	3
3	1	0	3	3	0	3	3	3
4	1	0	2	3	1	3	3	3
5	1	0	3	2	1	3	3	3
6	1	0	3	3	0	3	3	3

IRRELEVANT CLUSTER

Feature2
Feature5
Feature9
Feature12
Feature19
Feature20
Feature30

RELEVANT CLUSTER

Feature1
Feature3
Feature4
Feature6
Feature7
Feature8
Feature10
Feature11
Feature13
Feature14
Feature15
Feature16
Feature17
Feature18
Feature21
Feature22
Feature23
Feature24

Fig. 5 Segmentation process to test relevance in stage-3 of the feature engineering process.

Table 3. Comparisons of performance of classifiers before and after preprocessing

Ensemble Models	Before Preprocessing		After Preprocessing	
	Accuracy	ROC Curve	Accuracy	ROC Curve
Boosted Trees	78.1%	0.7246	96.9%	0.09516
Bagged Trees	68.6	0.7077	96.9%	0.09677
Subspace Discriminant	84.4%	0.7923	93.8%	0.0484
Subspace KNN	71.9%	0.7802	96.9%	0.1935
RU Boosted Trees	78.1%	0.7874	84.4%	0.09677

As shown in Table 3, the preprocessing has shown improvement in the accuracy of predictions of classifiers, as shown in Figure 6.

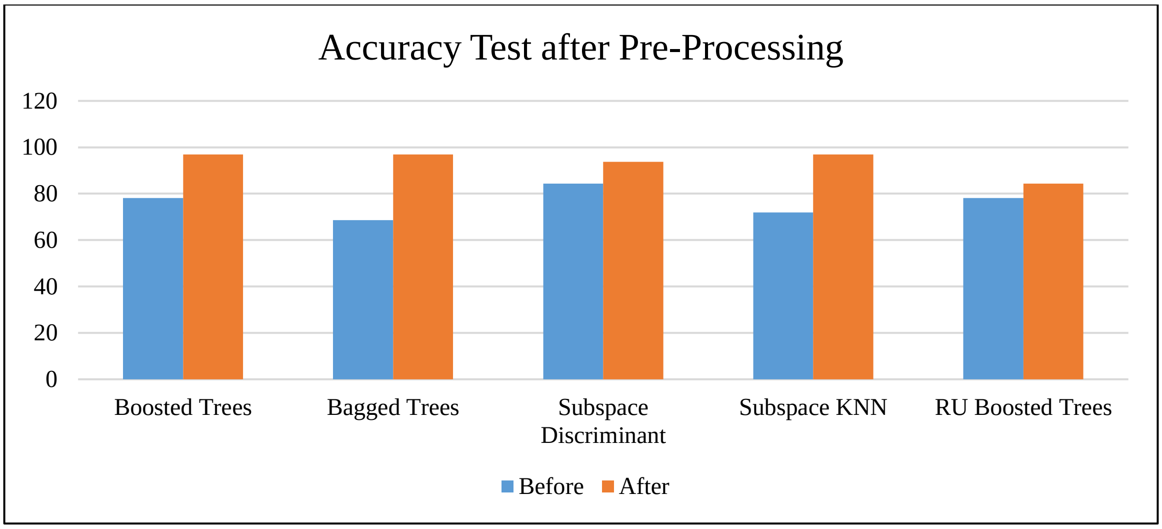


Fig. 6 The enhancement of accuracy after preprocessing of the dataset in three stages

As shown in Figure 6 the results of performance have augmented in all classifiers Boosted Trees (Before: 78.1%, After: 96.9%), Bagged Trees (Before: 68.6%, After: 96.9%), Subspace Discriminant (Before: 84.4%; After: 93.8%), Subspace KNN (Before: 71.9%; After: 96.9%), RU Boosted Trees (Before: 78.1%, After: 84.4%). Similarly, the preprocessing has shown improvement in the Region of Curve (ROC) of predictions of classifiers, as shown in Figure 7.

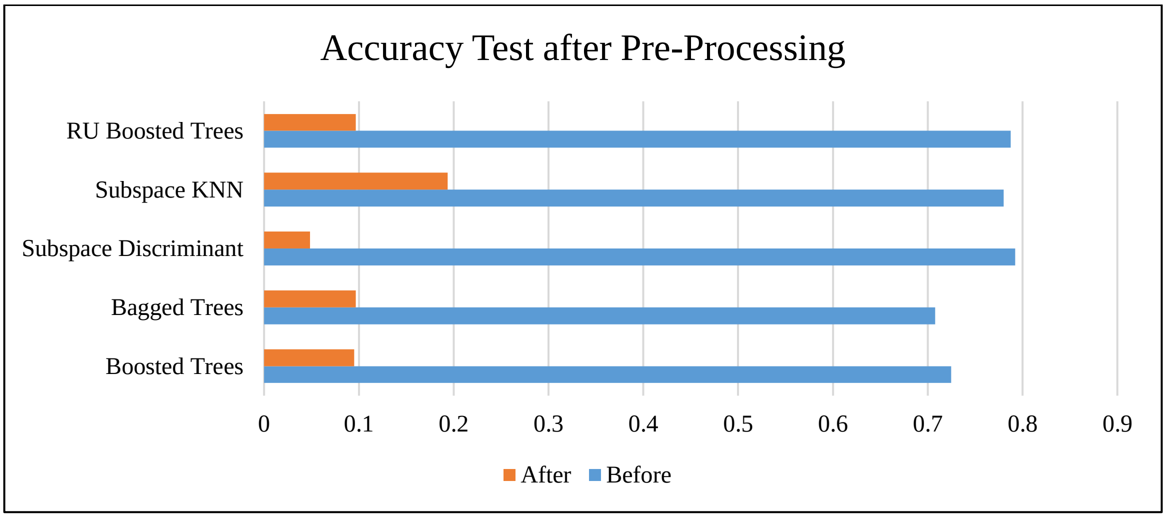


Fig. 6 The enhancement of the ROC curve after preprocessing of the dataset in three stages

As shown in Figure 7, the results of performance have augmented in all classifiers Boosted Trees (Before: 0.7246, After: 0.09516), Bagged Trees (Before: 0.7077, After: 0.09677), Subspace Discriminant (Before: 0.7923; After: 0.0484), Subspace KNN (Before: 0.7802; After: 0.1935), RU Boosted Trees (Before: 0.7874, After: 0.09677). The results show that the ROC, being an error value, has reduced considerably after preprocessing, and hence, the results were augmented after the experiment.

6. Conclusion

The outputs of this framework, a well-prepared and enhanced dataset, serve as a valuable resource for subsequent analysis, aiding researchers and healthcare professionals in making informed decisions regarding lung cancer diagnostics, prognostics, and treatments. In essence, the LC-PreProFE framework stands as a vital tool in the field of lung cancer research, enabling enhanced data utilization, facilitating advanced analytics, and ultimately contributing to the ongoing efforts to combat and manage lung cancer effectively. The test with the SNP Lung Cancer dataset has shown considerable augmentation in prediction; thereby, it is understood that by reducing the unwanted features of the lung cancer Multivariate dataset, the prediction with classifiers can be improved effectively.

References

[1] Chang Gu et al., "A Cloud-Based Deep Learning Model in Heterogeneous Data Integration System for Lung Cancer Detection in Medical Industry 4.0," *Journal of Industrial Information Integration*, vol. 30, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] V. Vasudha Rani, Smritilekha Das, and Tamal Kr. Kundu, “Risk Prediction Model for Lung Cancer Disease Using Machine Learning Techniques,” *Innovations in Computer Science and Engineering*, *Lecture Notes in Networks and Systems*, Singapore, vol. 385, pp. 417-425, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Roman Jaksik, and Jarosław Śmieja, “Prediction of Lung Cancer Survival Based on Multiomic Data,” *Intelligent Information and Database Systems*, *Lecture Notes in Computer Science*, Ho Chi Minh City, Vietnam, vol. 13758, pp. 116-127, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Negar Maleki, and Seyed Taghi Akhavan Niaki, “An Intelligent Algorithm for Lung Cancer Diagnosis Using Extracted Features from Computerized Tomography Images,” *Healthcare Analytics*, vol. 3, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Devendra K. Tayal et al., “A Novel Hybrid Approach for Dimensionality Reduction in Microarray Data,” *Proceedings of the International Conference on Intelligent Computing*, *Communication and Information Security*, *Algorithms for Intelligent Systems*, Singapore, pp. 213-226, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Rabia Musheer Aziz, “Application of Nature Inspired Soft Computing Techniques for Gene Selection: A Novel Frame Work for Classification of Cancer,” *Soft Computing*, vol. 26, pp. 12179-12196, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Muthuperumal Periyaperumal Ramkumar et al., “Deep Maxout Network for Lung Cancer Detection Using Optimization Algorithm in Smart Internet of Things,” *Concurrency and Computation: Practice and Experience*, vol. 34, no. 25, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] C. Venkatesan et al., “Efficient Machine Learning Technique for Tumor Classification Based on Gene Expression Data,” *2022 8th International Conference on Advanced Computing and Communication Systems*, Coimbatore, India, pp. 1982-1986, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Christopher J. Hanley et al., “Single-Cell Analysis Reveals Prognostic Fibroblast Subpopulations Linked to Molecular and Immunological Subtypes of Lung Cancer,” *Nature Communications*, vol. 14, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Kanchan Pradhan, Priyanka Chawla, and Sanyog Rawat, “A Deep Learning-Based Approach for Detection of Lung Cancer Using Self Adaptive Sea Lion Optimization Algorithm (SA-SLnO),” *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 12933-12947, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Guillaume Chassagnon et al., “Artificial Intelligence in Lung Cancer: Current Applications and Perspectives,” *Japanese Journal of Radiology*, vol. 41, pp. 235-244, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Kai Zhang et al., “Content-Based Image Retrieval with a Convolutional Siamese Neural Network: Distinguishing Lung Cancer and Tuberculosis in CT Images,” *Computers in Biology and Medicine*, vol. 140, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Na Sun et al., “A Novel 14-Gene Signature for Overall Survival in Lung Adenocarcinoma Based on the Bayesian Hierarchical Cox Proportional Hazards Model,” *Scientific Reports*, vol. 12, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Sergey P. Primakov et al., “Automated Detection and Segmentation of Non-Small Cell Lung Cancer Computed Tomography Images,” *Nature Communications*, vol. 13, pp. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Ebtasam Ahmad Siddiqui, Vijayshri Chaurasia, and Madhu Shandilya, “Detection and Classification of Lung Cancer Computed Tomography Images Using a Novel Improved Deep Belief Network with Gabor Filters,” *Chemometrics and Intelligent Laboratory Systems*, vol. 235, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Hariprasath Manoharan et al., “Aerial Separation and Receiver Arrangements on Identifying Lung Syndromes Using the Artificial Neural Network,” *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, pp. 1-8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Jovan Andjelkovic et al., “Sequential Machine Learning in Prediction of Common Cancers,” *Informatics in Medicine Unlocked*, vol. 30, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Nashat Alrefai, and Othman Ibrahim, “Optimized Feature Selection Method Using Particle Swarm Intelligence with Ensemble Learning for Cancer Classification Based on Microarray Datasets,” *Neural Computing and Applications*, vol. 34, pp. 13513-13528, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Ridhima Rani et al., “Big Data Dimensionality Reduction Techniques in IoT: Review, Applications and Open Research Challenges,” *Cluster Computing*, vol. 25, pp. 4027-4049, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] P.N. Senthil Prakash, and N. Rajkumar, “HSVNN: An Efficient Medical Data Classification Using Dimensionality Reduction Combined with Hybrid Support Vector Neural Network,” *The Journal of Supercomputing*, vol. 78, pp. 15439-15462, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Wei Yao et al., “Noninvasive Method for Predicting the Expression of Ki67 and Prognosis in Non-Small-Cell Lung Cancer Patients: Radiomics,” *Journal of Healthcare Engineering*, vol. 2022, no. 1, pp. 1-9, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Tarneem Elemam, and Mohamed Elshrkawey, “A Highly Discriminative Hybrid Feature Selection Algorithm for Cancer Diagnosis,” *The Scientific World Journal*, vol. 2022, no. 1, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Surbhi Gupta, and Yogesh Kumar, “Cancer Prognosis Using Artificial Intelligence-Based Techniques,” *SN Computer Science*, vol. 3, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Bikram Kar, and Bikash Kanti Sarkar, “A Hybrid Feature Reduction Approach for Medical Decision Support System,” *Mathematical Problems in Engineering*, vol. 2022, no. 1, pp. 1-20, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[25] Suli Liu, and Wu Yao, “Prediction of Lung Cancer Using Gene Expression and Deep Learning with KL Divergence Gene Selection,” *BMC Bioinformatics*, vol. 23, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[26] Saksham Gupta et al., “A Novel Transfer Learning-Based Model for Ultrasound Breast Cancer Image Classification,” *Computational Vision and Bio-Inspired Computing*, *Advances in Intelligent Systems and Computing*, Singapore, vol. 1439, pp. 511-523, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Moshood A. Hambali et al., “Feature Selection and Computational Optimization in High-Dimensional Microarray Cancer Datasets Via InfoGain-Modified Bat Algorithm,” *Multimedia Tools and Applications*, vol. 81, pp. 36505-36549, 2022. [CrossRef] [Google Scholar] [Publisher Link]