Egyptian Informatics Journal 23 (2022) 109–120

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Predicting the prevalence of lung cancer using feature transformation techniques

Zunaira Munawar [a], Fahad Ahmad [b,*], Saad Awadh Alanazi [c], Kottakkaran Sooppy Nisar [d], Madiha Khalid [e], Muhammad Anwar [f], Kashif Murtaza [e]

[a] Department of Computer Sciences, Kinnaird College for Women, Lahore, Punjab 54700, Pakistan
[b] Department of Basic Sciences, Deanship of Common First Year, Jouf University, Sakaka, Aljouf 72341, Saudi Arabia
[c] Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka, Aljouf 72341, Saudi Arabia
[d] Department of Mathematics, College Arts and Science, Prince Sattam Bin Abdulaziz University, Wadi Aldawaser 11991, Saudi Arabia
[e] Punjab University College of Information Technology, University of the Punjab, Lahore, Punjab 54700, Pakistan
[f] Department of Information Sciences, Division of Science and Technology, University of Education, Lahore, Punjab 54700, Pakistan

ARTICLE INFO

Article history:
Received 29 December 2021
Revised 21 June 2022
Accepted 30 August 2022
Available online 23 September 2022

Keywords:
Lung cancer
Carcinoma
Computerized tomography scan
Positron emission tomography scan
Machine learning
Dimensionality reduction
Feature transformation
Regression model

ABSTRACT

Healthcare sector is one of the most important sectors of any country as a big part of the country's economy is associated with it. The research is about to contribute to the health sector by minimizing the expenses of a lung cancer diagnosis. The study tries to devise an efficient method for initial screening of the patients with symptoms through their demographic and clinical data. The study seeks appropriate feature transformation techniques from dimensionality reduction techniques in combination with an apposite regression model that can perform this task robustly using the lung cancer dataset for early carcinoma diagnosis. To equip the health sector with state-of-the-art technology and for the betterment of humankind, the most beneficial tool of today is machine learning. Lungs play a vivacious role in the human body, oxygen is circulated in the body, and the air is taken from the atmosphere and sends it into the bloodstream. Several people decease every year because of lung carcinoma as lung cancer is normally identified at the latter phase due to lack of awareness. It stays unidentified because people have pneumonia often, which converts into lung cancer later. The projected research seeks to enable health professionals to the rationalization of primary diagnosis and treatment of lung carcinoma in developing countries. The proposed methodology has selected the optimized combination of regression-based machine learning technique and feature transformation technique for the available patterns based on demographic and clinical features of lung cancer patients. Based on the gathered results during training (RMSE = 0.1324, R2-Score = 0.7428) and testing (RMSE = 0.1273, R2-Score = 0.7405) it can be concluded that the fast independent component analysis and elastic net regression technique provided optimized results and outperform the other aspirant techniques.

1. Introduction

Lung diseases induce breathing system and pulmonary dysfunction and encompass diverse lung diseases. It includes pneumonia, tuberculosis, influenza, and lung cancer [1]. COVID-19 is a fatal lung ailment, and patients who survive are prone to asthma, pneumonia, and leukemia [2,3]. COVID-19 also causes respiratory failure death [4]. COVID-19 survivors have a reduced immune system, putting them at risk for pulmonary diseases. Multiple factors cause lung illnesses. Some are viral, bacterial, and fungal illnesses; others, like asthma, lung cancer, and mesothelioma, are environmental [5]. Bacterial and viral lung infections limit the

* Corresponding author.
E-mail addresses: drfahadahmadmian@ieee.org (F. Ahmad), sanazi@ju.edu.sa (S. Awadh Alanazi), n.sooppy@psau.edu.sa (K.S. Nisar), madiha.khalid@pucit.edu.pk (M. Khalid), anwar.muhammad@ue.edu.pk (M. Anwar), kashif.murtaza@pucit.edu.pk (K. Murtaza).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2022.08.002
1110-8665/

lungs' ability to hold air, which affects breathing. Air pollution, radon gas, asbestos, and some chemicals cause lung ailments [6].

Few lung cancer disorders are listed i.e., lung cancer, pneumonia, chronic pulmonary disease, asthma, etc. Cells proliferate abnormally in lung cancer. Sometimes it originates elsewhere and progresses to the lungs. Each cancer kind is treated differently [7]. The diagnosis of lung cancer was mentioned after mid-1700 as a severe disease that could take away life. In 1810 various physiognomies of lung cancer were identified. The American scientists of research center mentioned that lung cancer grew with the time. Lung carcinoma tumors increased and became the second most deadly cancer in 1940 [8,9]. It became the greatest common reason for demise in 2013, and it grew to 27 % of all other deaths.

A sputum cytology test can detect lung cancer cells. Patients who are having cough and producing sputum then observing the sputum under a microscope can detect cancerous cells. The diagnostic tool used earlier was bronchoscopy, by taking the samples of tissues, and the accuracy rate was 88 % [10]. In radial probe endobronchial ultrasound, a 360-degree image is obtained of the surrounding structure. Further, the targeted location is determined by inserting the probe into the traditional bronchoscopy for the exact lump position determination. A biopsy is again performed by the traditional biopsy technique [11]. The electromagnetic navigation system was used for tracking the working channel of bronchoscopy and to deploy the flexible tip of the probe by using a sensor attached to the probe. The sensor helps in creating a 3-dimensional position structure on computerized tomography (CT) scan and images obtained are used to determine the exact location of the lesions according to the sensor position with 68 %, accuracy [12].

Imaging tests include X-ray and CT scan, the difference between these two is that by doing X-ray abnormal mass can be detected from the image of the lungs whereas in CT scan small lesions in the lungs are detected, which may be neglected by X-ray. Furthermore, the stage of the cancer is determined by performing further required procedures, and those are CT, bones scan, magnetic resonance imaging (MRI), and positron emission tomography (PET). These scans help in deciding further treatment of cancer. The stage of the cancer is between 0 and 4; the lowest stage indicates that cancer is only in the lungs, and it is treatable with efficiency [13]. Where the last stage refers to that it has been spread in the other organs as well. Persons who are chain smokers tend to have small cell lung cancer and large cell carcinoma occurs in other people, and it includes non-small cell carcinoma [14].

The reasons for the spread, and types of lung cancer help in pinpointing the position of cancer. Pneumonia and lung cancer symptoms are quite similar like cough, chest pain, and shortness of breath. Furthermore, symptoms of pneumonia are rapid weight loss, high calcium levels, and more [15]. The indications which are not encompassed in lung carcinoma are chilling, shaking, and a large amount of sputum [16]. Research says, patients suffering from pneumonia often, tend to have lung carcinoma later. The causes of this cancer are mainly smoking, tobacco, and inhaling of chemicals in the surroundings as it causes cancerous cells to activate in the body [17,18]. Furthermore, cigarette smoke is hazardous as it contains known 60 carcinogens that include nitrosamines, and hydrocarbons, where p450 cytochrome enzymes activate polycyclic aromatic hydrocarbon. Lung cancer initiates by chronic DNA adduct formation in genes such as p53 [19,20].

Technology had contributed to the advancement of its use. Software-based solutions have been proposed to overcome the problem of today [21]. Identification of lung carcinoma issues can be overcome by the use of software-based solutions inspired by artificial neural networks (ANN) in machine learning (ML) [22–25]. Previously problem-solving was performed by conventional methods. ML-based methods provide the facility to resolve problems as in the human brain [26–29]. On the other hand, to improve the health sector of developing countries, ML methods like feature extraction and transformation techniques can be used to make the identification debauched and effective with high precision. In this era of technology, one more problem is the availability of redundant data, which can be handled by using feature transformation techniques that maps the high dimension features on low dimension space [30].

ML is the core subpart of artificial intelligence [31]. Without doing direct programming, ML, applications learn directly from patterns [32]. The unique thing is that when ML-based applications are exposed to new datasets, they learn to develop change and grow by themselves [33]. The learning process is iterative, which lets the ML learn from the existing data, which becomes its experience [34]. Without human interaction, computations are performed and results are produced in ML [35]. The lifeblood of all kinds of ML is data, and decisions based on data are definite causes of staying up with the competition or falling behind whereas ML is the way of going with the competition and staying ahead [36]. Moreover, ML would have a significant effect on living and the economy in the coming years as this is one of the most advanced technologies which would require less workforce and maximum automated work would be done [37].

The current research had been performed by using the algorithms of ML and feature transformation techniques that are core parts of contemporary ML. As we know, ML algorithms generate an output for the input which is given. A mathematical model is generated on the provided dataset as input and the patterns are formed by the model automatically. To increase the recital of the automatically formed model, feature transformation is performed. Feature transformation is a method in which data preprocessing is performed in a manner that unique patterns are identified [38]. It also helps in forming new features from the existing features. The new features may not have the same meaning as the original features, but they may have more power in different aspects [39]. In this study, diverse ML algorithms along with feature transformation techniques are performed on the lung cancer patient's dataset in order to identify the optimized model for its early diagnosis.

1.1. Problem Statement

The most difficult aspect of the planned research will be the early diagnosis of lung cancer, as it is typically diagnosed in its last stages, and its treatment at this stage is quite difficult [40]. The delay is caused in developing countries due to the communication gap between the medical professionals and the patients as many cannot afford the fees of private clinics [41]. The clinical diagnosis of lung cancer is an important topic in the modern era, and several early diagnostic systems for lung cancer have been proposed, mostly employing ML classification methods [42]. In these systems, the extraction of relevant characteristics posed the greatest challenge. To propose an effective approach for early diagnosis through the identification of robust features, the combination of feature transformation and ML algorithms has been used.

1.2. Aims and objective

In developing nations, the delay in cancer diagnosis is due to administrative, financial, and a shortage of available facilities, among others. In order to overcome the problem, the current research focuses on reducing the features of the available dataset of lung cancer patients through transformation and enhancing the performance of the diagnosis system through extensive learning. The purpose of this study is to detect lung cancer at an early stage. For this purpose, an ML algorithm comprised of a novel combination of regression and feature transformation techniques

would be developed. Dimensionality reduction of the presented dataset would be achieved using feature transformation techniques. In order to determine the optimal framework for the diagnostic system, a series of experiments would be conducted using root mean square error (RMSE) and R2-Score as performance metrics for the implemented schemes.

1.3. Novelty and contribution of the study

To detect disease in its earliest stages, numerous scientists have proposed various methods. Scientists have predominantly used image recognition for cancer detection. According to related research, diagnostic systems based on classification techniques have been proposed for the prevalent detection of cancer tumors. This study focuses on collecting sociodemographic, clinical, and laboratory investigation-based datasets of lung cancer patients and proposing an ML-based diagnosis system that includes a feature transformation technique and regression model for early detection. Optimization is quite challenging as in the recent world data is generating gigantically all the time and its simultaneous processing is quite big issue. So, the novelty and contribution of the presented research is identification of optimized combination contain regression technique and feature transformation technique.

1.4. Scope of the study

The purpose of this study is to propose a framework for the early diagnosis of lung cancer by transforming features. In addition, evaluate multiple dimensionality reduction techniques for feature transformation and select the optimal combination of regression model and feature transformation technique for early lung cancer diagnosis. In addition, the purpose of this study is to improve the accuracy of the model by using available datasets through enhanced pattern learning.

2. Materials and methods

In this research qualitative and quantitative analysis have been used to deploy the proposed framework. The model is functioning according to the stated flow in Fig. 1. Identified ML procedures are applied to train the model, which acts as an intelligent agent and will process the socio-demographic, clinical, and lab investigation-based dataset of lung cancer patients. In order to achieve appropriate diagnosis prediction accuracy and to avoid noise regression models are used with cross-validation techniques. As it is evident that lung cancer early diagnosis is a complex task, and the gathered dataset contains 48 features (Socio-Demo graphic = 11, Clinical = 23, Lab = 14) for that reason, the system is initially trained to reduce the dimensions of the dataset. The diverse feature transformation techniques from dimensionality reduction techniques are applied to extract governing features. Furthermore, the methodology is based on predictive modeling, the probabilistic process takes place to forecast the outcomes. It happens by using the inputs (Features) along with the output (Result) used. Regression techniques in this study are used for lung cancer prediction instead of classification; the reason is we are working in a numerical dataset, not categorical.

The study seeks to determine the best p and q for the proposed system. Where p denotes the regression model, and q denotes the feature transformation technique to contribute to the development of a system for early diagnosis of lung cancer. In the first module 'Dataset Collection' of the framework, the dataset is collected based on signs, symptoms, and findings. The identified dataset of lung cancer patients is processed and normalized between the range of 0–3 in the second module 'Parameters Detection and Characterization' and its characterization is done.

The third module is 'Regression-based Modelling', where regression-based seven techniques are applied for the prediction of the disease severity and stage. Furthermore, its training and testing error scores are recorded in the fourth module 'Planning and Monitoring' through a cross-validation process to minimize the error rate and avoid noise. Also, the probabilistic process from ML takes place, and learned the patterns according to the provided dataset. Ultimately, the minimal error rate regression technique is selected. The fifth module 'Dimensionality Reduction' tries to reduce the dimensionality of the preprocessed dataset and obtains optimized features by using feature transformation techniques. Different techniques of feature extraction are used to train the system and to identify the best technique which could be used to reduce the dimensions.

The sixth module is 'Decision Making' in this phase, according to the minimum RMSE, and the maximum R2-score of the applied techniques is identified. Then the best hybrid model based on regression and feature transformation techniques is selected. Unseen data is applied to the selected model, to validate the performance of the developed system. Finally, the trained system for lung carcinoma diagnosis is obtained.

The results of the selected model are shown after performing a series of experiments. Different techniques and models are used to validate the system for feature transformation of a complex sociodemographic, clinical, and lab investigation based dataset of lung cancer patients. The framework is the general representation of the experimentation that has been performed, and the results are shown step by step.

2.1. Optimized combination of regression model and feature transformation technique identification

Data collection is the initial step of this study, online source the National Cancer Institute is used to collect clinical data shown in Table 1 [43]. Dataset is of 3000 patients and the extracted features are 48 in number. The extracted features are: City, Area, Education, Marital Status, Occupation, Hobbies, Siblings, Age, Gender, Weight, Height, Family History, Hereditary Status, Smoking, Alcohol Consumption, Vomiting, Nausea, Temperature, Blood Pressure, Medicine Intake, Asthma, White Stools, Eye Color, Chilling, Cough, Weight Loss, Loss of Appetite, Back Pain, Obesity, Enlarged Liver, Fatigue/Weakness, Chest Pain, Gall Bladder Inflammation, Kidney Stone, Sputum Level, Sputum Color, Diabetes, Arthritis, Osteoporosis, Heart Disease/Attack, Hemoglobin Level, Liver Function Test, Forced Expiratory Volume in the First Second, Forced Vital Capacity, Hepatitis Type, Pneumonia, Bronchitis, Cancer Patient, Calcium Level, Result. After extraction of the pertinent features, normalization is done from 0 to 3 with relevant results as labels from 0 to 1 where each column represents an instance.

Then we will deploy ML techniques. Experiments were carried out on Lenovo Mobile Workstation equipped with Processor: 10th Generation Intel Core i7, Operating System: Windows 10 Pro 64, Memory: 16 GB DDR3, Hard Drive: 1 TB SSD, Graphics: NVIDIA RTX A3000. We have used Anaconda Prompt (Jupiter notebook) for the experimentation and results of our proposed scheme, and the language used in it is Python.

The dataset accessible to the trained system enables early detection of signs and symptoms. The target variable is unaffected, and its values are preserved. Due to the differences in the variables, the dataset is normalized between the scales 0 and 2. In this framework, the python built-in function has been used to split the dataset into training 70 % and testing 30 %. The model has been built on a training set, and performance has been evaluated on the testing set. Here, the values of y remain the same. This is known as the tar-

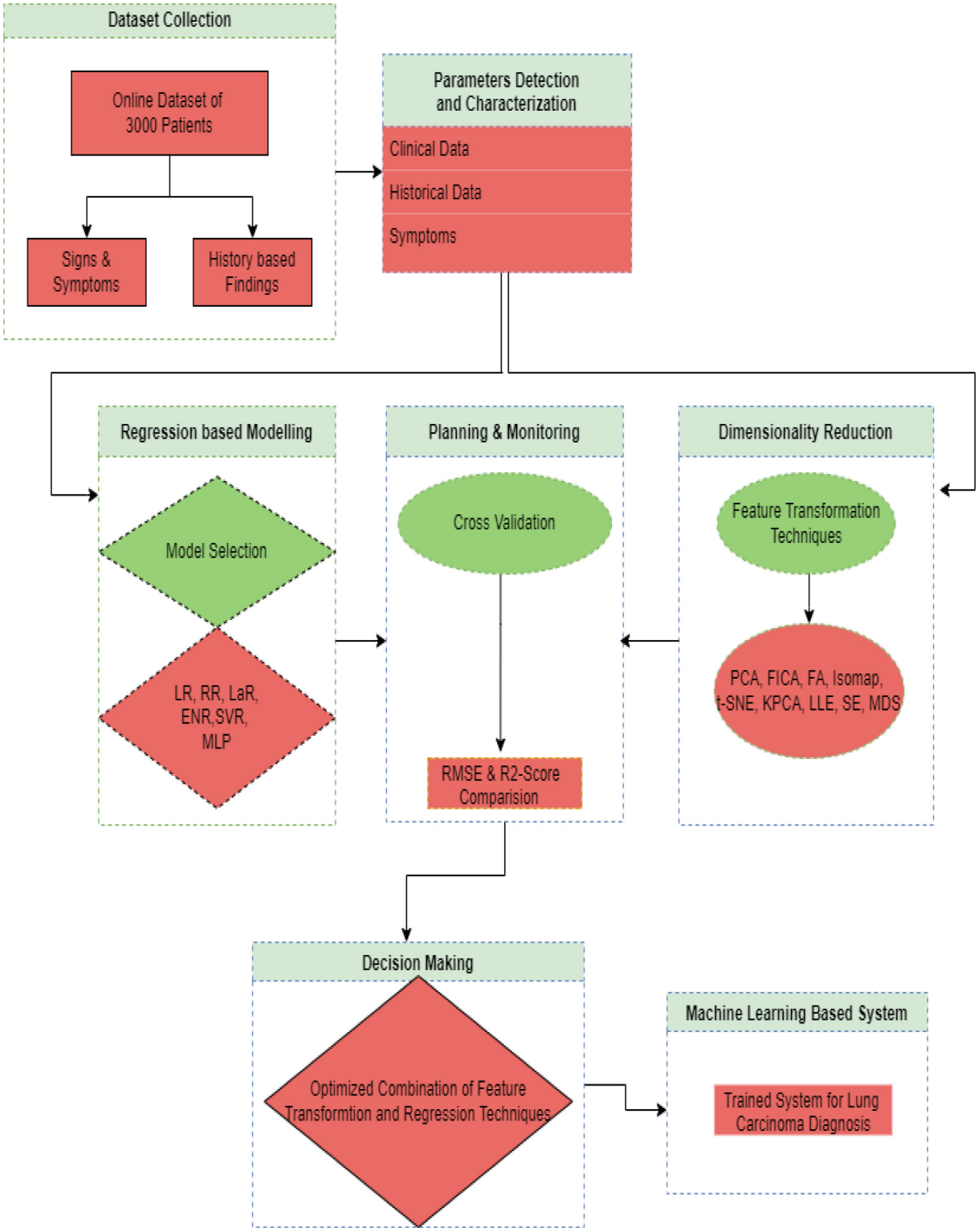

Fig. 1. Proposed Framework of Lung Cancer Early Diagnosis.

get value, where y is the result variable. Regression procedures are applied that are robust methods to determine the statistical relationship between the different independent variables and dependent variables.

The outcomes shown in Table 2 to Table 9 are predicted by regression models. The applied regression techniques are as follows: Linear Regression (LR), Ridge Regression (RR), Lasso Regression (LaR), Elastic Net Regression (ENR), Support Vector Regression (SVR), and Multilayer Perceptron (MLP). RMSE and R2-Score values are calculated against each regression technique performed, the lowest value of RMSE and highest value of R2-score show that model performance is best for the certain applied regression-based technique for the available dataset. Additionally, the closer together the numbers are, the lower the standard devi-

Table 1
Socio-Demographic, Clinical and Lab Investigation-based Dataset of Lung Cancer Patients.

Features	Values										
City	1	1	1	1	1	1	1	1	1	–	1
Area	1	0	1	0	0	0	1	0	0		0
Education	1	0	0	0	0	0	0	0	1		0
Marital Status	1	1	1	1	1	1	1	1	1		1
Occupation	0	1	1	1	2	2	0	0	2		2
Hobbies	0	3	3	1	1	3	3	2	2		3
Siblings	2	2	3	2	2	2	2	3	3		3
Age	1	1	1	0	2	1	1	1	2		0
Gender	0	0	0	0	0	0	0	0	0		0
Weight	1	2	1	1	2	1	0	1	2		2
Height	1	0	0	0	1	2	1	1	0		1
Family History	0	0	0	1	1	1	1	0	0		0
Hereditary Status	0	0	0	0	0	0	0	0	0		0
Smoking	1	1	1	1	1	1	1	0	1		1
Alcohol Consumption	0	1	0	1	0	1	1	1	1		1
Vomiting	0	0	0	0	0	0	0	0	0		0
Nausea	1	1	1	0	1	1	1	1	0		0
Temperature	1	1	1	1	1	1	1	1	1		1
Blood Pressure	1	1	1	1	1	1	1	1	1		1
Medicine Intake	0	0	0	0	0	0	0	0	0		0
Asthma	0	0	0	0	0	0	0	0	0		0
White Stools	0	0	0	0	0	0	0	0	0		0
Eye Color	0	0	0	0	0	0	0	0	0		0
Chilling	1	0	1	0	1	0	1	0	1		0
Cough	1	1	1	1	1	1	1	1	1		1
Weight Loss	1	1	1	1	1	1	1	1	1		1
Loss Of Appetite	1	1	1	1	1	1	1	1	1		1
Back Pain	1	1	1	1	1	1	1	1	1		1
Obesity	0	0	0	0	0	0	0	0	0		0
Enlarged Liver	0	1	0	0	0	0	0	0	0		0
Fatigue/Weakness	1	1	1	1	1	1	1	1	1		1
Chest Pain	1	1	1	1	1	1	1	1	1		1
Gall Bladder Inflammation	1	0	1	1	1	0	1	1	1		1
Kidney Stone	0	0	1	0	0	0	0	1	0		0
Sputum Level	1	0	0	0	0	0	0	0	0		0
Sputum Color	1	1	1	1	1	1	1	1	1		1
Diabetes	0	0	0	0	0	0	0	0	0		0
Arthritis	0	0	0	0	0	0	0	0	1		0
Osteoporosis	0	0	0	0	0	0	0	0	0		0
Heart Disease/Attack	0	0	0	0	0	0	0	0	0		0
Hemoglobin Level	0	1	1	0	1	1	1	2	1		1
Liver Function Test	0	0	0	0	0	0	0	0	0		0
Forced Expiratory Volume in the First Second	0	1	0	0	1	1	1	0	1		1
Forced Vital Capacity	0	1	1	0	1	1	1	0	0		1
Hepatitis Type	0	0	0	0	0	0	0	0	0		0
Pneumonia	1	1	1	1	1	1	1	1	1		1
Bronchitis	0	0	0	0	0	0	0	0	0		0
Cancer Patient	0	0	0	0	0	0	0	0	0		0
Calcium Level	0	0	0	0	0	0	0	0	0		0
Result	0.5	0.6	0.7	0.6	0.7	0.6	0.7	0.7	0.6		0.6

ation (STD) value is, which indicates a lesser likelihood of inaccuracy.

2.1.1. Linear Regression

An unpretentious form of regression is Linear Regression, which assumes that the prognosticators have a linear relationship with the target values, and Gaussian distribution takes place [44]. The Equation (1) of linear regression is as follows.

$$\mathbf{y} = \mathbf{a}_1\mathbf{x}_1 + \mathbf{a}_2\mathbf{x}_2 + \mathbf{a}_3\mathbf{x}_{3+} \cdots\cdots\cdots + \mathbf{b} \quad (1)$$

Table 2
Linear Regression RMSE and R2-Score for Training and Testing.

Performance Metric	Value
RMSE-Training	0.12173
RMSE-Testing	0.15530
R2-Score Training	0.77318
R2-Score Testing	0.67233

Where y is the target variable $x_1, x_2 \ldots$ are the features, $a_1, a_2 \ldots$ are the coefficients, and b is the parameter of the model.

2.1.2. Ridge Regression

Ridge Regression could be considered as the improved version of linear regression. In this model, complexity is reformed by the loss function the model [45]. This alteration is made by tallying a penalty of the parameters that are equivalent to the square of the magnitude of the coefficients as shown in Equation (2).

$$\textit{Loss Function} = \textit{Ordinary Least Squares} + \textit{Alpha} * \textit{Summation}(\textit{Squared Coefficient Values}) \quad (2)$$

In the above loss function, the value of the alpha needs to be selected by the developer as the low value of alpha leads to over-fitting and whereas a high alpha value can lead to under-fitting. Here, we have chosen the value of alpha = 0.01 and L2 as the regularization method, which adds the penalty term. Results are shown in Table 3.

Table 3
Ridge Regression RMSE And R2-Score for Training and Testing.

Performance Metric	Value
RMSE-Training	0.12172
RMSE-Testing	0.15522
R2-Score Training	0.77317
R2-Score Testing	0.67267

2.1.3. Lasso Regression

Another modified version of linear regression is Lasso Regression, which is also known as the least absolute shrinkage and selection operator [46]. The intricacy of the model is reduced by the loss function as shown by Equation (3).

$$Loss\ function = Ordinary\ Least\ Squares + Alpha * Summation\ (Absolute\ Values\ of\ The\ Coefficients'\ Magnitude) \quad (3)$$

In the above equation, the L1 regularization term is used. It stands for the least absolute shrinkage and selection operator. Results are shown in Table 4.

2.1.4. Elastic Net Regression

The combination of ridge and lasso regression forms Elastic Net Regression and works by reprimanding the model [47]. In scikit-learn, the elastic net class is used to construct the elastic net regression model. This model is used most commonly with L1 and L2 regularization terms as shown by Equation (4) as follows:

$$\frac{\sum_{i-1}^{n}(Y_i - X_i)^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\beta j + \alpha\sum_{j=1}^{m}\beta j\right) \quad (4)$$

Here, chooses the value of alpha between 0 and 1. Where, if alpha = 0, it will correspond to the ridge regularization term, and if we select alpha = 1, it will correspond to the lasso regularization term. Hence, here the value of alpha has been chosen between 0 and 1. Results are shown in Table 5.

2.1.5. Support Vector Regression (SVR)

Linear and non-linear regression problems are solved by using Support Vector Regression. In this margin violations are limited, and it fits as many instances as possible [48]. Results are shown in Table 6.

Table 4
Lasso Regression RMSE And R2-Score for Training and Testing.

Performance Metric	Value
RMSE-Training	0.13836
RMSE-Testing	0.14647
R2-Score Training	0.70693
R2-Score Testing	0.70855

Table 5
Elastic Net Regression RMSE and R2-Score for Training and Testing.

Performance Metric	Value
RMSE-Training	0.13384
RMSE-Testing	0.14888
R2-Score Training	0.72579
R2-Score Testing	0.69886

Table 6
Support Vector Regression, RMSE, and R2-Score for Training and Testing.

Performance Metric	Value
RMSE-Training	0.12897
RMSE-Testing	0.16035
R2-Score Training	0.74892
R2-Score Testing	0.65070

2.1.6. Multi-Layer Perceptron (Neural network based Model)

Multi-layer Perceptron (MLP) works for both classification problems and regression problems. This algorithm is supervised, it learns a function f(.): R^m -> R^o, by training on a dataset. The number of inputs is denoted by m, and the number of outputs is denoted by o, for the given dataset. It consists of hidden layers, and these layers are present between the input and output layers [49,50]. These layers are non-linear. It is different from logistic regression; MLP works like a decision tree [51]. Results are shown in Table 7.

The results shown above are from deployed regression models for lung cancer detection. Lasso regression model RMSE and R2-Score are better. Furthermore, cross-validation techniques will be performed.

Due to the variations and amount of data ML-based different models are emerging. Innovative ML procedures have made it easy to handle critical datasets. However, at the time of training, the dataset can influence the performance of the algorithm. The complexity of the algorithm also differs with the size of the dataset which has a great impact on the training and testing procedures and results.

2.2. Cross-Validation techniques

Cross-validation is one of the most important phases applied in the current framework. It is used to avoid noise in the system. The model is selected by using a cross-validation technique in every step to avoid any noise and validate the performance of the hidden patterns by using the latest learned capabilities. It helps in testing unseen data on the system trained by ML techniques. In this study, k-folds and leave-one-out cross-validation techniques have been used to achieve accuracy [52,53]. Results of both techniques are shown in Table 8 and Table 9.

2.3. Dimensionality reduction techniques

The process of reducing the total number of features in the feature set is called dimensionality reduction. This process is done by

Table 7
Multi-Layer Perceptron Regression RMSE and R2-Score for Training and Testing.

Performance Metric	Value
RMSE-Training	0.10200
RMSE-Testing	0.25760
R2-Score Training	0.84830
R2-Score Testing	0.00800

Table 8
K-Fold Cross Validation Technique.

Polynomial Degree	RMSE	Standard Deviation (STD)
1	0.06902544721622078	0.09547423534738518
2	0.06902544721622076	0.09547423534738513
3	0.06902544721622070	0.09547423534738511
4	**0.06902544721622067**	**0.09547423534738508**

Table 9
Leave One Out Cross Validation Technique.

Leave One Out	RMSE
Linear Regression	0.02369423522813748
Ridge Regression	**0.02310380958238347**
Lasso Regression	0.06902544721622078
Elastic Net Regression	0.06902544721622076

using strategies like feature selection and feature extraction or feature transformation. The feature selection technique only considers the features that have great influence and the rest are not selected [54]. In the feature extraction or feature transformation technique, new features are formed from the present features. The dataset of high dimensionality is reduced to a small dimensional space.

Advantages of dimensionality reduction techniques are:

- Reduces storage space and compresses data.
- It takes less time for computations.
- Redundancy is removed by it if any exists.

Disadvantages of dimensionality reduction are:

- Data loss is possible in a small amount.
- The principal component analysis (PCA) found linear correlations which are not required in some conditions.
- Its results are not accurate sometimes, in a condition where mean and covariance are not enough to define datasets.
- PCA follows some thumb rules, in order to know how many principal components to keep in practice.

The researchers gave importance to the feature transformation technique for dimensionality reduction and according to them, original features are replaced by the functions of the provided features, this phenomenon is known as feature transformation. The function could be of an individual feature or a group of features, where this function itself is a random variable. According to the scientists, feature transformation exactly acts like changing the bases in calculus, and the bases of feature space were changed in feature transformation. It had helped in working and debauched speed. [55].

The feature transformation techniques and their results are as follows:

- Principal Component Analysis (PCA)
- Fast Independent Component Analysis (fastICA)
- Factor Analysis (FA)
- Isomap
- *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE)
- Kernel Principal Component Analysis (KPCA)
- Locally Linear Embedding (LLE)
- Multidimensional Scaling (MDS)
- Spectral Embedding (SE)

2.3.1. Principal Component Analysis

Principal Component Analysis (PCA) is a technique for unsupervised learning datasets. This technique of feature transformation works by mapping the data space of higher dimensional to lower-dimensional space provided that variance should be maximum [56,57].

As the values of RMSE are comparatively lower than others during training and testing using LR model along with PCA. Similarly, the values of R2-Score are comparatively higher than others during training and testing using LR model along with PCA. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using LR and PCA as highlighted in Table 10.

2.3.2. Fast Independent Component Analysis

The Fast Independent Component Analysis (fastICA) is also one of the techniques of dimensionality reduction [58]. For blind source separation issues, the fastICA technique is the most popular approach since it is computationally effective and uses less memory than other blind source separation algorithms.

As the values of RMSE are comparatively lower than others during training and testing using ENR model along with fastICA. Similarly, the values of R2-Score are comparatively higher than others during training and testing using ENR model along with fastICA. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using ENR and fastICA as highlighted in Table 11.

2.3.3. Factor Analysis

FA is one of the feature transformation techniques in which essential factors and covert variables from a set of observed features are identified by using the data analysis method. It helps in data elucidations by reducing the number of variables. It extracts maximum common variance from all features and puts them into

Table 10
Principal Component Analysis along with Regression Techniques.

Principal Component Analysis	RMSE Training	RMSE Testing	R2-Score Training	R2-Score-Testing
Linear Regression	**0.13984**	**0.15087**	**0.71494**	**0.65963**
Ridge Regression	0.13980	0.15097	0.71494	0.65964
Lasso Regression	0.14035	0.15117	0.71280	0.65829
Elastic Net Regression	0.13999	0.15096	0.71430	0.65922
Support Vector Regression	0.11340	0.16890	0.81250	0.57342
Multi-Layer Perceptron Regression	0.15310	0.17420	0.65850	0.54620

Table 11
Fast Independent Component Analysis along with Regression Techniques.

Fast Independent Component Analysis	RMSE Training	RMSE Testing	R2_Score Training	R2_Score Testing
Linear Regression	0.118932	0.159469	0.793813	0.6197414
Ridge Regression	0.118938	0.159247	0.793794	0.6207969
Lasso Regression	0.140182	0.154754	0.713550	0.6418900
Elastic Net Regression	**0.1344920**	**0.148361**	**0.736334**	**0.6708707**
Support Vector Regression	0.1082705	0.187509	0.829124	0.4742599
Multi-Layer Perceptron Regression	0.1966000	0.217600	0.436700	0.2919000

a normal range [59]. Factor analysis is widely utilized in advertising, market research, finance, operation research, and many more. Market researchers use factor analysis to identify price-sensitive customers, identify brand features that influence consumer choice, and help in understanding channel selection criteria for the distribution channel. The Equation (5) of factor analysis is shown below.

$$\boldsymbol{Y_i} = \boldsymbol{\beta_{i0}} + \boldsymbol{\beta_{i1}}\boldsymbol{F}_1 + \boldsymbol{\beta_{i2}}\boldsymbol{F}_2 + (1)\boldsymbol{e_i} \tag{5}$$

In this method number of variables are investigated and to see whether several variables of concentration X_1, X_2, ... are linearly related to an unobservable smaller number of factors F_1, F_2,..... F_k.

As the values of RMSE are comparatively lower than others during training and testing using ENR model along with FA. Similarly, the values of R2-Score are comparatively higher than others during training and testing using ENR model along with FA. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using ENR and FA as highlighted in Table 12.

2.3.4. Isomap

Isomap is a method that works well for the non-linear dataset. It is one of the dimensionality reduction techniques which uses a Geodesic distance approach for the multivariate data points. Some methods work for the linear dataset, which reduces the dimensions based on Euclidean distances [60].

As the values of RMSE are comparatively lower than others during training and testing using ENR model along with Isomap. Similarly, the values of R2-Score are comparatively higher than others during training and testing using ENR model along with Isomap. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using ENR and Isomap as highlighted in Table 13.

2.3.5. t-Distributed Stochastic Neighbor Embedding

The *t*-Distributed Stochastic Neighbor Embedding is a dimensionality reduction technique that picks similar samples that have a high likelihood, and different points are having meager chances of being picked. The technique represents a high-dimensional dataset in a low-dimensional space of two or three dimensions so that it can be visualized conveniently [61]. Then, *t*-SNE defines a similar distribution for the points in the low-dimensional embedding and finally, minimizes the Kullback–Leibler divergence between the two distributions concerning the locations of the points in the embedding. In Equation (6) take a partial derivative of the cost of every point to get the update of each point direction.

$$\frac{\delta C}{\delta y_i} = 2\sum_j \left(p_{j|i-q_{j|i}} + p_{i|j\_}q_{i|j}\right)(y_{i-}y_j) \tag{6}$$

As the values of RMSE are comparatively lower than others during training and testing using LR model along with *t*-SNE. Similarly, the values of R2-Score are comparatively higher than others during training and testing using LR model along with *t*-SNE. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using LR and *t*-SNE as highlighted in Table 14.

2.3.6. Kernel Principal Component Analysis

Kernel Principal Component Analysis (KPCA) is a feature transformation technique from dimensionality reduction techniques for the non-linear dataset. Each point is reconstructed from its neighbor distance [62]. For a large dataset, it is difficult to represent graphically.

As the values of RMSE are comparatively lower than others during training and testing using LR model along with KPCA. Similarly, the values of R2-Score are comparatively higher than others during training and testing using LR model along with KPCA. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using LR and KPCA as highlighted in Table 15.

2.3.7. Locally Linear Embedding

Local Linear Embedding (LLE) works well for non-linear datasets its algorithm works in the following manner: initially, get the neighbors of each data point, Xi, then, compute the weights for the data points, which would best reconstruct it by minimizing the cost, finally computes the vectors Yi, reconstructed best by the weights, and quadratic form is minimized [63].

As the values of RMSE are comparatively lower than others during training and testing using ENR model along with LLE. Similarly, the values of R2-Score are comparatively higher than others during training and testing using ENR model along with LLE. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using ENR and LLE as highlighted in Table 16.

Table 12
Factor Analysis along with Regression Techniques.

Factor Analysis	RMSE Training	RMSE Testing	R2_Score Training	R2_Score Testing
Linear Regression	0.118932	0.159469	0.793813	0.619741
Ridge Regression	0.118933	0.159433	0.793813	0.619913
Lasso Regression	0.131173	0.160486	0.749186	0.614874
Elastic Net Regression	**0.126549**	**0.150226**	**0.766557**	**0.662545**
Support Vector Regression	0.099413	0.184146	0.855939	0.492947
Multi-Layer Perceptron Regression	0.103900	0.240800	0.842600	0.132800

Table 13
Isomap along with Regression Techniques.

Isomap	RMSE Training	RMSE Testing	R2_Score Training	R2_Score Testing
Linear Regression	0.118932	0.159469	0.793813	0.619741
Ridge Regression	0.118932	0.1594329	0.793813	0.619913
Lasso Regression	0.131173	0.1604862	0.749187	0.61487
Elastic Net Regression	**0.1265493**	**0.1502257**	**0.766557**	**0.662550**
Support Vector Regression	0.0994128	0.1841464	0.855939	0.492947
Multi-Layer Perceptron Regression	0.1074000	0.2655000	0.832000	−0.054200

Table 14
t-Distributed Stochastic Neighbor Embedding along with Regression Techniques.

t-Distributed Stochastic Neighbor Embedding	RMSE Training	RMSE Testing	R2_Score Training	R2_Score Testing
Linear Regression	0.118932	0.159460	0.793813	0.619740
Ridge Regression	0.118930	0.159433	0.793813	0.619913
Lasso Regression	**0.131176**	**0.149680**	**0.749175**	**0.664988**
Elastic Net Regression	0.126550	0.150210	0.766550	0.662300
Support Vector Regression	0.099400	0.184150	0.855900	0.4929400
Multi-Layer Perceptron Regressor	0.088200	0.206400	0.886500	0.362900

Table 15
Kernel Principal Component Analysis along with Regression Techniques.

Kernel Principal Component Analysis	RMSE Training	RMSE Testing	R2_Score Training	R2_Score Testing
Linear Regression	0.118932	0.159500	0.793813	0.619740
Ridge Regression	0.118933	0.159433	0.7938125	0.619913
Lasso Regression	**0.131170**	**0.149680**	**0.7491752**	**0.664988**
Elastic Net Regression	0.126540	0.150220	0.766550	0.662300
Support Vector Regression	0.099400	0.184140	0.855900	0.492940
Multi-Layer Perceptron Regressor	0.103900	0.237800	0.842600	0.154400

Table 16
Locally Linear Embedding along with Regression Techniques.

Locally Linear Embedding	RMSE Training	RMSE Testing	R2-Score Training	R2-Score Testing
Linear Regression	0.118900	0.159400	0.793800	0.619740
Ridge Regression	0.118932	0.159430	0.794800	0.619910
Lasso Regression	0.131173	0.160480	0.749100	0.614800
Elastic Net Regression	**0.126540**	**0.150230**	**0.766550**	**0.662500**
Support Vector Regression	0.099410	0.184140	0.855930	0.492900
Multi-Layer Perceptron Regressor	0.102100	0.237100	0.84820	0.159200

2.3.8. Multidimensional Scaling

Multidimensional Scaling (MDS) is used to represent similarities in the dataset, and it analyzes non-linear data. A similarity matrix or any kind of distance can be observed by using MDS; it is its beauty. These similarities can represent people's ratings of similarities between objects, the percent agreement between judgments, the number of times a subject fails to discriminate between stimuli, etc [64].

As the values of RMSE are comparatively lower than others during training and testing using ENR model along with MDS. Similarly, the values of R2-Score are comparatively higher than others during training and testing using ENR model along with MDS. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using ENR and MDS as highlighted in Table 17.

2.3.9. Spectral Embedding

It is one of the feature transformation techniques used to create groups of similar data points in a large dataset. The values of the elements in a matrix show the similarity and dissimilarity of the elements. If the value of the element is 0, it means they are similar where the high value shows the dissimilarity [65]. It is mostly used for linear datasets.

As the values of RMSE are comparatively lower than others during training and testing using ENR model along with SE. Similarly, the values of R2-Score are comparatively higher than others during training and testing using ENR model along with SE. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using ENR and SE as highlighted in Table 18.

3. Results

It is observed from the performed experiments results from that out of all deployed regression models lasso regression outperforms the other. The results have been shown as combined in the below Table 19 where the values of RMSE and R2_Score are close for all the regression models. The better result has been chosen based on a slight difference.

To ensure the performance of the system, six regression techniques have been used. The models have been tuned for 3000 lung cancer patient datasets. In the field of medical science, the effi-

Table 17
Multidimensional Scaling along with Regression Techniques.

Multidimensional Scaling	RMSE Training	RMSE Testing	R2_Score Training	R2_Score Testing
Linear Regression	0.118932	0.159469	0.793813	0.619741
Ridge Regression	0.118933	0.159433	0.793813	0.619913
Lasso Regression	0.131173	0.160486	0.749186	0.614874
Elastic Net Regression	**0.126549**	**0.150226**	**0.766557**	**0.662545**
Support Vector Regression	0.099412	0.184146	0.855939	0.492946
Multi-Layer Perceptron Regressor	0.101600	0.252200	0.849500	0.049200

Table 18
Spectral Embedding along with Regression Techniques.

Spectral Embedding	RMSE Training	RMSE Testing	R2_Score Training	R2_Score Testing
Linear Regression	0.118932	0.159470	0.794900	0.619800
Ridge Regression	0.118932	0.159433	0.793813	0.619900
Lasso Regression	0.131100	0.160500	0.749100	0.614900
Elastic Net Regression	**0.126540**	**0.150220**	**0.766550**	**0.662500**
Support Vector Regression	0.099500	0.184150	0.855900	0.492950
Random Forest Regressor	0.055600	0.156200	0.955000	0.635050
Multi-Layer Perceptron Regressor	0.106200	0.246600	0.835500	0.090900

Table 19
Regression Models' Results.

Regression Models	RMSE Training	RMSE Testing	R2-Score Training	R2-Score Testing
Linear Regression	0.121726	0.155300	0.773180	0.672330
Ridge Regression	0.121700	0.155220	0.773100	0.672600
Lasso Regression	**0.138300**	**0.146470**	**0.706900**	**0.708500**
Elastic Net Regression	0.133840	0.148880	0.725700	0.698800
Support Vector Regression	0.128000	0.160300	0.748900	0.650700
Multi-Layer Perceptron Regressor	0.102000	0.257600	0.848300	0.008000

ciency of the system cannot be compromised. The system has been trained and tested for the dataset of lung cancer patients and its influencing features have been identified. New features are produced from the existing features by using unique feature transformation techniques that are being used with the regression models [66]. Furthermore, different transformation techniques' results are compared to select the best hybrid model, and after this phase, the unseen data has been run on the selected hybrid model to check the overall efficiency of the proposed model. In Table 20 best combination of feature transformation technique and regression model has been highlighted that presents the best results. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using LaR as highlighted in Table 19.

The highlighted row shows the best p and q, where p is a transformation technique, and q is the regression model. According to the results fastICA, along with ENR, hybrid model surpassed the other models. The unseen dataset has been simulated on the trained system by using the Jupiter prompt and python language. By using the selected regression model, i.e., ENR and feature transformation technique, i.e., fastICA unseen dataset has been imitated on the model. The combined effect is basically a tradeoff between RMSE and R2-Score and represents their best combination using ENR and fastICA as highlighted in Table 19.

Table 21 and Table 22 are showing the results of the training and testing phases. During training, an unseen dataset is not included. According to the results, the system efficiency is 74 %,

Table 20
Feature Transformation Techniques along with Regression Techniques.

Feature Transformation Techniques-Regression Models	RMSE Training	RMSE Testing	R2-Score Training	R2-Score Testing
Principal Component Analysis-Linear Regression	0.13984000	0.150871000	0.7149400	0.659630000
Fast Independent Component Analysis-Elastic Net Regression	**0.13449200**	**0.148361000**	**0.7363338**	**0.670870662**
Factor Analysis-Elastic Net Regression	0.126549327	0.150225725	0.766556894	0.66254515
Multidimensional Scaling- Elastic Net Regression	0.126549327	0.150225725	0.766556894	0.66254515
Spectral Embedding-Elastic Net Regression	0.126540000	0.150220000	0.766550000	0.66250000
Kernel Principal Component Analysis-Lasso Regression	0.131170000	0.149680000	0.749175193	0.66498800
t- Distributed Stochastic Neighbor Embedding- Lasso Regression	0.131176000	0.149680000	0.749175193	0.66498800
Locally Linear Embedding-Elastic Net Regression	0.126540000	0.150230000	0.766550000	0.66250000
Isomap-Elastic Net Regression	0.1265493270	0.150225725	0.766556894	0.66254515

Table 21
Final Training Results.

Selected Feature Transformation Technique-Regression Model	RMSE Training	R2-Score Training
Fast Independent Component Analysis-Elastic Net Regression	**0.13240123134195655**	**0.742809928631397**

Table 22
Final Testing Results of Unseen Data.

Selected Feature Transformation Technique- Regression Model	RMSE Testing for Unseen Dataset	R2-Score Testing for Unseen Dataset
Fast Independent Component Analysis-Elastic Net Regression	**0.1273386167500566**	**0.7405118957283781**

Table 23
Features Importance Score.

Scores
0.18197167
−0.0584853
0.00470465
−0.015664
0.09607283
0.09280731
0.06586555
−0.03095821
0.06810358
−0.01436014
−0.19050213
0.07343331
−0.36827166
−0.07827317
0.08787992

and the error rate is 13 % approximately during the training and testing phases. But the room for improvement is still there.

In the below Table 23, the scores of the features are shown after transformation. The transformed features' importance is known based on values. The features having a value higher than 0.01 are known to be the more critical features according to the trained system specifications. Features having negative values show that these features are not influential.

4. Discussion

The efficient system has been proposed by combining robust feature transformation and regression techniques. The system has been trained for the complex dataset of 3000 patients. The system aims to diagnose lung cancer at its early stage. Many regression techniques have been used to see the efficiency of the system, i.e., LR, RR, LaR, ENR, SVR, and MLP. By comparing the results of all the regression techniques then the optimized model was selected. Diverse feature transformation techniques have been applied i.e., PCA, fastICA, FA, Isomap, *t*-SNE, KPCA, LLE, MDS, and SE. Each transformation technique was performed along with all the regression models to get the best result. After performing all the identified steps, the hybrid of fastICA and ENR was selected. In the comparative studies, researchers have proposed diagnostic systems for other diseases named Alzheimer and liver tumor detection, in those systems they had used ML image recognition techniques. They had used diverse classification and regression techniques, and the accuracy rate of those systems was approximately around 75 %–80 % [67]. Whereas the importance of the feature transformation technique is that it saves time by reducing the dimension of the dataset into small space. To achieve accuracy original features are not disturbed. On the trained machine, the unseen dataset is simulated to check the accuracy of the system. The results have shown the efficiency of the trained system. Researchers had also used the artificial neural network technique to predict the presence of lung cancer in the human body, according to the applied technology prediction accuracy was approximately 80 % [68]. Where in the current research the techniques which are applied give the accuracy of 74 % for the clinical dataset.

5. Conclusion

This research has proposed a combination of the feature transformation technique and regression model after performing a series of experiments. The combination of fastICA and ENR gives an accuracy of 74 % and an error rate of 13 % only after running unseen data on the trained system. The results can be slightly different if any other dataset would be used. The trained system aims to contribute to the early diagnosis of lung cancer by using ML techniques special for resource constraint countries from the third world. Hence, the combination of fastICA and ENR gives the best result for the early diagnosis of lung cancer. In the proposed framework, the accuracy rate is not very high, but it lies in a reasonable range. It shows that more work can be done to improve the performance. Regression is used in this research instead of classification due to the nature of the study and dataset that is numerical, not categorical. To make developing countries' health sectors better, ML techniques can play a vital role. Our future direction based on this study could be stage identification after obtaining appropriate features through feature transformation in dimensionality reduction. If a patient's condition is critical, radiological examinations such as computed tomography, magnetic resonance imaging, or positron emission tomography might be recommended. ML algorithms can also be used to recommend these radiological tests. After that, deep learning algorithms might be applied to these images to help pinpoint the tumor's location in the specific organ. Deep learning algorithms will also be used to identify cancer cells and their spread to other organs.

6. Data availability statement

The datasets used for this study is publicly available.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

Any organization did not financially support this present work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Thanks to our families and colleagues who supported us morally.

References

[1] L, S., T. , and S. Hodi, Five-Year Survival and Correlates Among Patients,With Advanced Melanoma, Renal Cell Carcinoma,or Non–Small Cell Lung Cancer Treated With Nivolumab. JAMA Oncol, 2019: p. 1516.
[2] Ahmad F et al. Prediction of COVID-19 cases using machine learning for effective public health management. Comput Mater Continua 2021:2265–82.
[3] Guite, H., COVID-19: What happens inside the body. Retrieved from Medical News Today: https://www. medicalnewstoday.com/articles/co vid-19-what-happens-inside-the-body, 2020.
[4] Aguiar D et al. Inside the lungs of COVID-19 disease. Int J Legal Med 2020;134 (4):1271–4.
[5] Denis F et al. Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. JAMA 2019;321 (3):306–7.
[6] Goodwin JS et al. Use of the shared decision-making visit for lung cancer screening among Medicare enrollees. JAMA Int Med 2019;179(5):716–8.
[7] Retico A, Fantacci ME. The potential contribution of artificial intelligence to dose reduction in diagnostic imaging of lung cancer. J Med Artif Intell 2019:1–5.

[8] Sang, J., M.S. Alam, and H. Xiang. Automated detection and classification for early stage lung cancer on CT images using deep learning. in Pattern Recognition and Tracking XXX. 2019. SPIE.
[9] Mehmood, M., et al., Improved Colorization and Classification of Intracranial Tumor Expanse in MRI Images via Hybrid Scheme of Pix2Pix-cGANs and NASNet-Large. Journal of King Saud University-Computer and Information Sciences, 2022.
[10] Madan B, Panchal A, Chavan D. Lung cancer detection using deep learning. 2nd international Conference on Advances in Science & Technology (ICAST), 2019.
[11] Arbour KC, Riely GJ. Systemic therapy for locally advanced and metastatic non–small cell lung cancer: a review. JAMA 2019;322(8):764–74.
[12] Asuntha A, Srinivasan A. Deep learning for lung Cancer detection and classification. Multimedia Tools Appl 2020;79(11):7731–62.
[13] Rehman MZ et al. Lungs cancer nodules detection from CT scan images with convolutional neural networks. International Conference on Soft Computing and Data Mining. Springer; 2020.
[14] Liang, W., et al., Society for Translational Medicine consensus on postoperative management of EGFR-mutant lung cancer (2019 edition). Translational lung cancer research, 2019. 8(6): p. 1163.
[15] Bakulski KM et al. DNA methylation signature of smoking in lung cancer is enriched for exposure signatures in newborn and adult blood. Sci Rep 2019;9 (1):1–13.
[16] Witschi H. Tobacco smoke–induced lung cancer in animals—a challenge to toxicology (?). Int J Toxicol 2007;26(4):339–44.
[17] Cruz CSD, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. Clin Chest Med 2011;32(4):605–44.
[18] Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. RSNA 2018:1–6.
[19] MacKinnon AC, Kopatz J, Sethi T. The molecular and cellular biology of lung cancer: identifying novel therapeutic strategies. Br Med Bull 2010:47–61.
[20] Crosbie PA et al. Yorkshire Lung Screening Trial (YLST): protocol for a randomised controlled trial to evaluate invitation to community-based low-dose CT screening for lung cancer versus usual care in a targeted population at risk. BMJ open 2020;10(9):e037075.
[21] Malhotra J et al. Risk factors for lung cancer worldwide. Eur Respir J 2016:889–902.
[22] Shanid M, Anitha A. Lung Cancer Detection From CT Images Using SALP-Elephant Optimization-Based Deep Learning. Biomed Eng: Appl Basis Commun 2020:1–3.
[23] Chakraborty D, Elzarka H. Advanced machine learning techniques for building performance simulation: a comparative analysis. J Build Perform Simul 2019:193–207.
[24] Mehmood M et al. Machine learning enabled early detection of breast cancer by structural analysis of mammograms. Comput Mater Continua 2021;67 (1):641–57.
[25] Ngiam KY, Khor W. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019;20(5):e262–73.
[26] Shabbir, M., et al., Cognitively managed multi-level authentication for security using Fuzzy Logic based Quantum Key Distribution. Journal of King Saud University-Computer and Information Sciences, 2022.
[27] Ammarah Cheema MT, Hafiz A, Khan MM, Ahmad FH, Anwar M. Prevention Techniques against Distributed Denial of Service Attacks in Heterogeneous: A Systematic Review. Security and Communication. Networks 2022. 2022..
[28] Hasan T et al. Edge Caching in Fog-Based Sensor Networks through Deep Learning-Associated Quantum Computing Framework. Comput Intell Neurosci 2022;2022.
[29] Yanes N et al. Fuzzy Logic Based Prospects Identification System for Foreign Language Learning Through Serious Games. IEEE Access 2021;9:63173–87.
[30] . Ahmad, A.K., A. Jafar, and K. Aljoumaa Customer churn prediction in telecom using machine learning in big data platform. Springer, journal of big data, 2019: p. 1-6..
[31] Abd Ghani MK, Mohammed MA. In: Decision-level fusion scheme for nasopharyngeal carcinoma identification using machine learning techniques. Springer; 2018. p. 625–38.
[32] Hunt GJ, Dane MA. Automatic Transformation and Integration to Improve Visualization and Discovery of Latent Effects in Imaging Data. J Comput Graph Sta 2019:1–20.
[33] Wu Z et al. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc Natl Acad Sci 2019;116(18):8852–8.
[34] Finlayson SG et al. Adversarial attacks on medical machine learning. Science 2019;363(6433):1287–9.
[35] Chapaneri R, Shah S. A Comprehensive Survey of Machine Learning-Based Network Intrusion Detection. Springer; 2018. p. 345–56.
[36] Zorc JJ, Chamberlain JM, Bajaj L. Machine learning at the clinical bedside—the ghost in the machine. JAMA Pediatr 2019;173(7):622–4.
[37] Fu, G.-S., et al., Machine learning for medical imaging. 2019, Hindawi.
[38] Aggarwal, T., A. Furqan, and K. Kalra. Feature extraction and LDA based classification of lung nodules in chest CT scan images. in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2015. IEEE.
[39] Hirakawa, T., K. Wararatpanya, and Y. Kuroki. Image recognition using multi-layer sparse feature extraction with ADMM. in International Workshop on Advanced Image Technology (IWAIT) 2019. 2019. SPIE.
[40] DeSantis, C.E., et al., Cancer treatment and survivorship statistics, 2014. CA: a cancer journal for clinicians, 2014. 64(4): p. 252-271.
[41] Mathew AR et al. Life-Course Smoking Trajectories and Risk for Emphysema in Middle Age: The CARDIA Lung Study. American Journal of Critical Care Medicine 2019:1–4.
[42] Pradhan K, Chawla P. Medical Internet of things using machine learning algorithms for lung cancer detection. J Manage Analy 2020;7(4):591–623.
[43] Institute, N.C., Lung Datasets. 2019: Cancer Data Access System.
[44] Toğaçar M, Ergen B, Sertkaya ME. Zatürre Hastalığının Derin Öğrenme Modeli ile Tespiti. Firat University. J Eng 2019;31(1).
[45] Mehmood M et al. Systematic Framework to Predict Early-Stage Liver Carcinoma Using Hybrid of Feature Selection Techniques and Regression Techniques. Complexity 2022;2022.
[46] McEligot AJ et al. Logistic LASSO regression for dietary intakes and breast cancer. Nutrients 2020;12(9):2652.
[47] Zhao H et al. Associations of prenatal heavy metals exposure with placental characteristics and birth weight in Hangzhou Birth Cohort: Multi-pollutant models based on elastic net regression. Sci Total Environ 2020;742:140613.
[48] Başaran E et al. Chronic Tympanic Membrane Diagnosis Based on Deep Convolutional Neural Network. IEEE; 2019.
[49] Alanazi SA et al. Estimation of Organizational Competitiveness by a Hybrid of One-Dimensional Convolutional Neural Networks and Self-Organizing Maps Using Physiological Signals for Emotional Analysis of Employees. Sensors 2021;21(11):3760.
[50] Aslam B et al. Ozone depletion identification in stratosphere through faster region-based convolutional neural network. CMC-Comput Mater Continua 2021;68(2):2159–78.
[51] Shahzadi S et al. Machine learning empowered security management and quality of service provision in SDN-NFV environment. Computers Materials Continua 2021;66(3):2723–49.
[52] Rafało M. Cross validation methods: analysis based on diagnostics of thyroid cancer metastasis. ICT Express 2021.
[53] Sultan LR et al. Machine Learning to Improve Breast Cancer Diagnosis by Multimodal Ultrasound. IEEE; 2018.
[54] Basavegowda HS, Dagnew G. Deep learning approach for microarray cancer data classification. CAAI Trans Intell Technol 2020;5(1):22–33.
[55] Ashcroft DM. The Essentials of Data Analytics and Machine Learning. Data Sceince Center of Excellence 2016:1–17.
[56] Toğaçar M, Ergen B. Biyomedikal Görüntülerde Derin Öğrenme ile Mevcut Yöntemlerin Kıyaslanması. Fırat Üniversitesi Mühendislik Bilimleri Dergisi 2019;31(1):109–21.
[57] Garcia-Larsen V et al. Dietary patterns derived from principal component analysis (PCA) and risk of colorectal cancer: a systematic review and meta-analysis. Eur J Clin Nutr 2019;73(3):366–86.
[58] Spurek P, Tabor J, Śmieja M. Fast independent component analysis algorithm with a simple closed-form solution. Knowl-Based Syst 2018;161:26–34.
[59] Sun S et al. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol 2019;20(1):1–21.
[60] Liu H et al. Dimensionality reduction for identification of hepatic tumor samples based on terahertz time-domain spectroscopy. IEEE Trans Terahertz Sci Technol 2018;8(3):271–7.
[61] Zhou H, Wang F, Tao P. t-Distributed stochastic neighbor embedding method with the least information loss for macromolecular simulations. J Chem Theory Comput 2018;14(11):5499–510.
[62] Mushtaq, Z., et al. Performance analysis of supervised classifiers using PCA based techniques on breast cancer. in 2019 international conference on engineering and emerging technologies (ICEET). 2019. IEEE.
[63] Jiao C-N et al. Hyper-graph regularized constrained NMF for selecting differentially expressed genes and tumor classification. IEEE J Biomed Health Inf 2020;24(10):3002–11.
[64] Zhuang H et al. Dysbiosis of the gut microbiome in lung cancer. Front Cell Infect Microbiol 2019;9:112.
[65] Guarracino MR et al. Classification of cancer cell death with spectral dimensionality reduction and generalized eigenvalues. Artif Intell Med 2011;53(2):119–25.
[66] Fernán AJ, Iribarnea L, Correl A. A recommender system for component-based applications using machine learning techniques. Elsevier; 2019. p. 68–84.
[67] Geng-Shen-Fu ML. for Medical Imaging. J Healthcare Eng 2019:1–2.
[68] Nasser IM, Naser SSA. Lung Cancer Detection Using Artificial Neural Network. Int J Eng Inf Syst 2019:17–23.