Journal of Information Systems Engineering and Management
2025, 10(3)

e-ISSN: 2468-4376

https://www.jisem-journal.com/ Research Article

Leveraging Enhanced PSO and Proving Random Forest's Dominance for Prediction of Lung Cancer Severity

Sanjeev Prakashrao Kaulgud[1*], Rama Krishna K[2], Mrutyunjaya M S[3], Abhinandan Shirahatti[4], Murthy D H R[5], Afroz Pasha[6]

[1] Department of Artificial Intelligence & Machine Learning, New Horizon College of Engineering, Bengaluru, Karnataka, India.
[2] Department of Artificial Intelligence & Machine Learning, Impact College Of Engineering & Applied Sciences, Bengaluru, Karnataka, India.
[3] Department of CSE(Data Science), R L Jalappa Institute of Technology, Doddaballapura, Karnataka, India.
[4] Department of Computer Science Engineering, Kolhapur Institute of Technology's College of Engineering, Kolhapur, Maharashtra India.
[5] Department of CSE(Cyber Security), R L Jalappa Institute of Technology, Doddaballapura, Karnataka, India.
[6] School of Computer Science Engineering, Presidency University, Bengaluru, Karnataka, India.
*Corresponding author Email: sanjeev.kaulgud@gmail.com

ARTICLE INFO

Received: 08 Nov 2024

Revised: 29 Dec 2024

Accepted: 23 Jan 2025

ABSTRACT

One of the most fatal malignancies in the world is lung cancer. Usually, it starts in the cells lining the airways of the lung tissues. Close to 85% of lung cancer cases are caused by smoking, and hence making it the primary reason of the disease. However, lung cancer can also strike non-smokers, or passive smokers, for a variety of reasons, including genetics, asbestos exposure, radon gas exposure, and second hand smoke. Frequent signs and symptoms include a chronic cough, chest discomfort, sudden weight loss, dyspnea, etc. The high death rate linked to delayed diagnosis of lung cancer makes the need for machine learning algorithms to anticipate the disease more and more obvious. The lung cancer usually shows no symptoms (asymptomatic) till it progresses to an advanced stage. Hence, early identification and diagnosis become critical to increase the survival rates. Huge datasets, such as genetic profiles of patients, their histories, and medical pictures may be processed and analysed using the machine learning to find the patterns and risk factors that primitive approaches might miss. Timely identification made possible by these models allows for prompt treatments and more individualized treatment regimens. In this work, Enhanced Particle Swarm Optimization (PSO) is used to Pre-Process the data. It draws inspiration from collective behavior of swarm and uses Interval-Newton method. We used machine learning algorithms such as Random Forest, K-Nearest Neighbors, Multinomial Logistic Regression, and Support Vector machines in this paper to analyse the available data. We found that Random Forest performed better than other methods with 97% for Low, 94% for Medium, and 98% for High severity.

Keywords: Lung Cancer, Machine Learning, Random Forest, Enhanced PSO, Support Vector Machine, Multinomial Logistic Regression, K-Nearest Neighbors, Prediction.

1. INTRODUCTION

In medical terms, 'cancer' describes a set of diseases characterized by out-of-control growth and spread of abnormal cells in the body. These cells can grow into tumors, they can invade close tissues and they can travel to other parts of the body through either the bloodstream or the lymphatic system. The main categories of cancer are classified based on the tissue or organ in which they develop. For instance, breast, lung, prostate, cervical and leukemia to name a few. Malignancies are a broad category of diseases characterized by the uncontrolled growth and spread of abnormal cells throughout the body. These aberrant cells have the ability to form masses, penetrate nearby structures, and spread via lymphatic and circulatory channels to distant parts of the body. Healthy cells follow cycle of 1) growth, 2) division, and 3) death. But the cancerous cells do not follow the cycle.

They multiply uncontrollably and avoid the body's all natural defense mechanism. Based on the location and type of the abnormal growth, cancer can develop in nearly any tissue or organ. Leukemia, breast cancer, and lung cancer are some of the well-known cancers. Genetic alterations, environmental exposures (such as radiation, tobacco smoke etc.), lifestyle factors (such as alcohol consumption, physical inactivity etc.), and specific infections are some of the multifactorial causes of cancer.

Despite many advancements in research and health treatments, cancer remains one of the most important reasons of death across the world. Early & Timely recognition and treatments are important in improving survival rates. This can be achieved only when cancer is diagnosed at an earlier stage. Current researchers have explored new ways to detect and treat cancer to prevent its global impact.

In the year 2024, the cancer will still be a major health issue in India, with the new cases recording over 1.5 million annually. Delayed cancer identification leads to a higher mortality rate. The study says that breast cancer is the most common type of cancer among the female population. In the men's population, it is oral cancer is a common one. Enhancing healthcare facilities and raising cancer awareness among people will help reduce the mortality rate. Cancer cases are expected to surpass 1.7 million annually by 2030. It is imperative to prioritize improving prevention tactics, early detection, and availability of efficient treatment to address this growing issue.

This is one of the most common diseases caused by a combination of genetic, environmental, and lifestyle factors; treatment includes surgery, radiation therapy, chemotherapy, immunotherapy, or targeted therapies. Advances in detection, together with early detection, have dramatically improved the outcome for many patients diagnosed with cancer.

According to recent data [1], in the year 2022 roughly 1,461,427 cases of cancer were diagnosed in India. It corresponds to an approximate 100 incidence rate for every 1,00,000 people. According to projections, approximately one in nine Indians will receive a cancer diagnosis at some point in their lives. Male patients were more likely to have pulmonary cancers than female patients, with mammary carcinomas being the most common cancer found in female patients. Lymphoid leukemia became the most common type in pediatric cases (ages 0–14), impacting 292 percent of male children and 242 percent of female children.

Disability Adjusted Life Years (DALYs) are predicted to rise from 26.7 million in 2021 to 29.8 million in 2025, with a greater impact in the northern and north-eastern regions of the country than in other regions. The incidence of cancer in India is predicted to increase by 13% between 2020 and 2025. Cancer is placed at second place with 18%, while cardiovascular disease is the leading with 63% for cause of death among non-communicable diseases. According to data from the National NCD Monitoring Survey (2017–2018), 32.8 percent of adults used alcohol and tobacco, 41.3 percent were physically sedentary, 96.4 percent did not eat enough fruits and vegetables, and the average daily salt intake was 8 grams. These variables highlight the urgent public health concern that India's growing cancer incidence presents.

The sample statistics as given in the report of American Cancer Society [2], are given below in Table 1.
Smoking is the leading factor, contributing to about 80 to 90% of all cases of lung cancer. This emphasizes how important it is to maintain ongoing public health initiatives to lower the prevalence of smoking and promote early screening, especially for populations who are more vulnerable.

According to the recent reports, lung cancer is a serious health issue throughout the world. In terms of the number of deaths from cancer as of 2024, lung cancer continues to lead the list in both the United States and India. This is valid for both genders. The majority of individuals had advanced diagnoses, which drastically reduces the number of available treatments and chances of survival. The late discovery of the illness is mostly to blame for this tendency.

2. LITERATURE REVIEW

C. Anil Kumar et. al [3] presented in their article that, due to the complex structure of cancer cells and also the variety of cancer types, the medical field is posed with a challenge of prediction of lung cancer. Delays in treating lung cancer greatly increase the risk of mortality, while early detection and intervention can lead to better outcomes. they demonstrated that both time and financial resources can be saved by usage of Support Vector Machine (SVM) and informing the patient in time for further action to be taken. The SVM-based method was rigorously evaluated and produced encouraging results, indicating its potential utility as a tool for oncologists to identify lung cancer cases.

Maurya, S. P. et. al [4] proposed a system that makes use of the patient's lifestyle and presents a low-risk status of cancer, predicts cancer that uses symptom identification making it more efficient. This system can help the specialists to recommend a suitable treatment strategy which can be tailored to the person's cancer risk profile. But, achieving higher accuracy is important in lung cancer prediction. After 310 cases were examined to find positive examples based on gender, each positive case was compared on the basis of gender for a variety of features. The research included a comprehensive evaluation of twelve different machine learning algorithms. Among these, the K-nearest neighbor and Bernoulli Naïve Bayes models emerged as top performers, achieving accuracy rates of 92.86% and 91.07%, respectively.

Venkatesh et. al [5] gave a study that focused on lung cancer, one of the predominant cancer types, and analyzed survival predictions using data from the SEER program. The original SEER dataset, which included 149 attributes and 1,000 samples, was reduced to 24 attributes following initial preprocessing. The study evaluated the effectiveness of Bagging and AdaBoost ensemble methods, utilizing three base learners: Decision Tree, K-Nearest Neighbor and Neural Network, for predicting lung cancer survival. The study also demonstrated that the bootstrap aggregating technique improved the performance of individual models, yielding accuracy scores of 0.982 for Decision Tree with AdaBoost, 0.951 for K-Nearest Neighbor with AdaBoost, and 0.931 for Neural Network with both Bagging and AdaBoost.

TABLE 1: American Cancer Society Statistics

	Estimated New Cases			Estimated Deaths		
Type \ Gender	**Both**	**Male**	**Female**	**Both**	**Male**	**Female**
Leukemia	62770	36450	26320	23670	13640	10030
Urinary Sys-	169360	118330	51030	32350	22360	9990
Lung & Bron-	**234580**	**116310**	**118270**	**125070**	**65790**	**59280**
Breast	313510	2790	310720	42780	530	42250
Other types	1220920	755200	465720	387850	220480	167370
All Cases	**2001140**	**1029080**	**972060**	**611720**	**322800**	**288920**

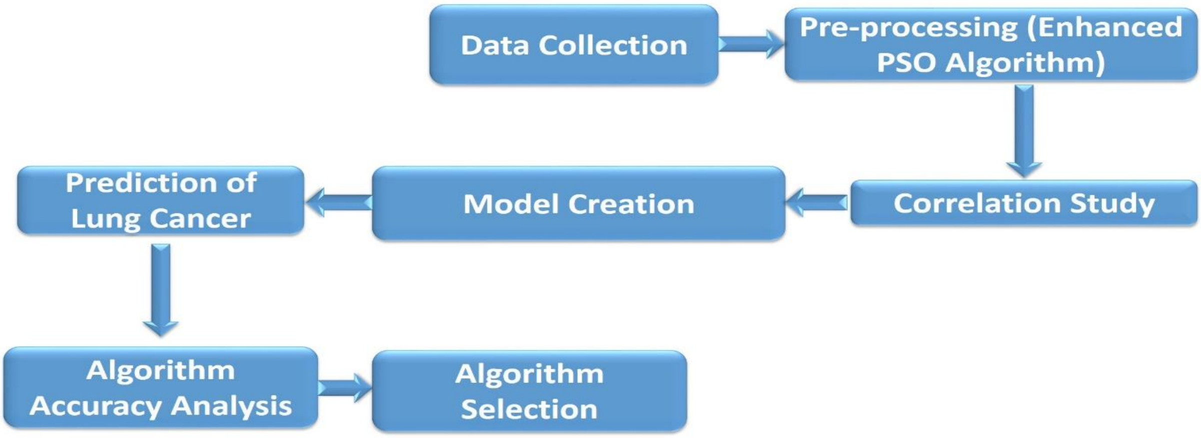
Data Collection
Pre-processing (Enhanced PSO Algorithm)
Correlation Study
Model Creation
Prediction of Lung Cancer
Algorithm Accuracy Analysis
Algorithm Selection

Figure 1: Methodology

Janee Alam et. al [6] proposed an algorithm and applied to lung cancer detection and prediction, yielding quite promising results when compared with other techniques employed. Such textural features are extracted from the separated ROIs and classified by the SVM. Thus, the algorithm will identify if there are some tumor cells in the input image and predict the possible development. With this proposed algorithm, the precision result arrives approximately at 97% for the cancer identification and around 87% for the cancer prediction. The system is likely to assist doctors in identifying whether or not the lung is dangerous and non-carcinogenic. The precision of the system may be developed on a large set of images and can be integrated with genetic algorithms and deep neural networks..

Syed Saba Raoof et. al [7] presented a comprehensive survey on lung cancer, including its causes, symptoms, and mortality rates both in India and globally. It explores the application of machine learning techniques in healthcare, particularly in cancer prognosis and detection. Researchers have predominantly used supervised learning techniques and classification algorithms to develop cancer prediction systems that yield accurate results. The paper also highlights the importance of deep learning in healthcare, emphasizing its potential to enhance the accuracy of lung cancer identification and prediction. By employing deep learning techniques, the prediction and diagnosis systems for lung cancer could be further refined.

3. METHODOLOGY

The methodology for predicting lung cancer starts with a thorough data pre-processing, addressing missing data, encoding categorical variables, and applying normalization or standardization. Correlation analysis plays a crucial role in understanding the relationships between various factors in lung cancer research, helping to identify potential risk factors, predict outcomes, and guide treatment strategies. The Methodology adopted in this work is represented in Figure 1. This structured approach ensures the development of a dependable and effective lung cancer prediction model.

3.1 Data Collection

The dataset contains 1000 samples and 24 attributes related to lung cancer analysis, encompassing various factors that may contribute to the disease's development or progression. These attributes include demographic information such as Age and Gender, environmental factors like Air Pollution and Dust Allergy, lifestyle choices such as Alcohol use, work-related risks captured by Occupational Hazards, genetic predisposition represented by Genetic Risk, and pre-existing conditions like Chronic Lung Disease. The target attribute, 'Level', likely indicates the severity or stage of lung cancer, serving as the dependent variable for predictive modelling or classification tasks. This comprehensive set of features allows researchers and data scientists to explore the complex interplay of multiple risk factors in lung cancer development and potentially identify key predictors for early detection or risk assessment. The details of all the attributes are given in Table 2.

Summary of data

Identification of Patients: The ID for each patient will be identified in the column of Patient Id.

Demographics: A column like Age and Gender can be present to gather basic demographic information about every patient. It is evident that Gender is encoded as an integer - 1 denotes male, and 2-female.

Risk and Lifestyle Factors: Air Pollution, Alcohol use, Dust Allergy, Occupational Hazards, and Genetic Risk factors are included to capture environment and lifestyle risks. Each of these factors is probably measured on a scale, possibly capturing the intensity of exposure or the level of risk.

Health Conditions: Columns are included for chronic Lung Disease, Obesity, and Smoking which outline health conditions that could cause or result from the factors listed above. These too are again represented numerically

Symptoms: Several columns of symptoms include Chest Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, amongst hundreds of others. The symptom is captured with a scale which may be indicative of intensity or duration.

Outcome Level: This is the column which indicated the seriousness or stage that has been reached by the disease in possibly such labels as "Low," "Medium," and "High".

3.2 Pre-processing using Enhanced PSO algorithm

Particle Swarm Optimization (PSO) is a unique computational technique that draws inspiration from the collective behaviour of swarms, setting it apart from other evolutionary algorithms. To begin, a swarm of a specified size is established. Each individual within this swarm, referred to as a particle, possesses distinct characteristics.

Initially, the population of particles is initialized with random position and velocity values. These positions are possible solutions in the search space. An unbiased function value for every particle is then assessed based on its present position coordinates.

As the algorithm proceeds, each particle preserves its personal best position called Pbest. This is simply the top solution that a specific particle has found up to this point in time. In parallel with this, information maintained by the swarm is on the global best position, referred to as Gbest, which is the best solution found by any particle in the entire swarm. It enables a PSO search to balance the experiences of individual particles with the collective experience within the swarm to converge on optimal solutions.

The particle i velocity is indicated as Eq. (1), and the particle i location is signified as Eq. (2). In addition, every particle i has historically best position, which is specified as $h_i = [h_{i1}, h_{i2,} .. h_{id}]$. The particle's best position is identified by using the position of neighbourhood particles that is specified as $n_i = [n_{i1}, \ldots\ldots n_{id}]$. The vectors $X_0$and V_i are updated randomly by utilizing the equations (1) and (2). In this PSO variant, several key parameters are introduced to guide the particles' movement. The inertia weight, denoted as ω, influences the particle's tendency to maintain its current trajectory. Two acceleration coefficients, c1 and c2, control the impact of personal and global best positions on the particle's movement.

Additionally, r1 and r2 are random values generated within the range [0, 1] for each dimension of the search space, adding a stochastic element to the algorithm.

This particular PSO variant with an interval-Newton method is focused on the proper degree of diversity to be maintained by the swarm through adaptive changes in the search space and improvement in the quality of local search capabilities. Interval-Newton is a very suitable method for solving a system of non-linear equations, where parameters may take on interval values.

Definition In this context we define a continuous function 'f' that is continuously differentiable. Here, Equation (3) as defined in the attached document formally represents this function. Integration of interval-Newton method with PSO aims at enhancing the capacity of an algorithm in better navigation over complex solution spaces for convergence to optima.

To solve the equation$F(X) = 0$, select the starting point. $X_0 = [X_0^1, X_0^2, \ldots X_0^d]$. Then, apply an iterative formula in Eq. (3) as represented in Eq. (4). Where, J is represented as the Jacobian matrix of F, and then introduce the interval–newton method $N(X) = J^{-1}(X_n)F(X_n)$ in Eq. (4), which is defined in Eq. (5) and (6). In this formulation, several key components are defined to describe the particle's movement in the PSO algorithm. The initial velocity is denoted as v0, while x0 represents the starting position. For any given particle i, its current velocity is expressed as vi(t), and its previous position is indicated by xi(t-1).

TABLE 2: Overview of all input features in the lung cancer study dataset

Attribute	Type	Values
Age	Age	Numeric Value
Gender	Gender	M [Male], F [Female]
Air Pollution, Alcohol use, Balanced Diet, Smoking, Passive Smoker	Habit	Scaled Value*
Dust Allergy, Obesity, Occupational Hazards, Chest Pain, Genetic Risk, Chronic Lung Disease, Coughing of Blood, Fatigue, Weight , Shortness of Breath Loss, Wheezing, Swallowing Difficulty, Clubbing of Finger Nails, Frequent Cold, Dry Cough, Snoring	Symptom	Scaled Value*
Level	Symptom	Low, Medium, High
***Scale of 1 [Least] to 10 [Highest]**		

$$V_i = [V_{i1}, V_{i2,} \ldots . V_{id}] \quad (1)$$

$$X_i = [X_{i1}, X_{i2,} \ldots . X_{id}] \quad (2)$$

$$F(X^1, X^2, \ldots X^d) = \begin{bmatrix} f_1(X^1, X^2, \ldots X^d) \\ \vdots \\ f_d(X^1, X^2, \ldots X^d) \end{bmatrix} = 0 \quad (3)$$

$$X_{n+1} = X_n - J^{-1}(X_n)F(X_n), \qquad n = 0,1,2 \dots \tag{4}$$

$$X_{n+1} = X_n - N(X),\ n = 0,1,2 \dots \tag{5}$$

$$X_{n+1} = X_n + V_{n+1} \tag{6}$$

$$V_{n+1} = wV_n + BN(X_n) \tag{7}$$

The interval-Newton method integrates with the PSO algorithm by updating the particle's position. This update is achieved by adding the current velocity to the particle's position. The calculation of the current velocity for particle i involves both an inertia weight and acceleration constants. This velocity update mechanism is mathematically expressed in Equation (7) of the original text.

This approach combines the exploratory nature of PSO with the precision of the interval-Newton method. By incorporating these elements, the algorithm aims to balance global exploration and local exploitation, potentially leading to more efficient convergence and improved solution quality in complex optimization problems. Here, w is denoted as inertia weight, and B is indicated as acceleration constant.

3.3 Correlation Study

One of the ways through which correlations might result in the disease is with complex relationships that might occur in lung cancer datasets. Through correlation analysis, possible risk factors will be found, and perhaps hidden patterns will emerge to improve on accuracy in predictive models. This includes the study of the strength and direction that exist in relationships between age, smoking history, genetic traits, environmental exposures, and clinical indicators. Strong correlations may indicate interesting insights. Like, it may be that the longer an individual smokes, the worse the cancer is, or certain genetic markers indicate more effective treatments. Furthermore, knowledge of these correlations enables one to figure out which features to use in machine learning models for improved accuracy and interpretation. Figure 2 displays the correlation of all attributes. It will be extremely helpful in lung cancer re-search, which may lead to better diagnostic methods, personalized treatments, and more effective prevention strategies.

3.4 Development of a Model for Lung Cancer Detection

There are many machine learning algorithms that can be used for this, relative strengths and weaknesses for each. One of the simplest yet effective algorithms for multiclass classification is Multinomial Logistic Regression.
Another powerful combination of multiple decision trees, known as Random Forest Classifier, that enhances the predictiveness of accuracy and also deals very well with large feature sets, often leading to models that are more robust and better suited for more complicated data. Gaussian Naive Bayes : It is simple and fast but useful in high-dimensional spaces, even though each feature is assumed to be independent with the actual data. KNN is a non-parametric method based on a simple heuristic: simple but computationally expensive on large datasets. SVM can work rather well in high-dimensional spaces by finding the best hyperplane that separates different classes of severity.

The matrix suggests some key variables on lifestyle and environmental factors, among which are "air pollution," "smoking," "alcohol use," "occupational hazards," and "balanced diet." Several of these variables show strong correlations. For example, "air pollution" and "occupational hazards" had a positive correlation of 0.75 indicating that exposure to occupational hazards would be higher in areas having increased air pollution. Analogously, "using alcohol" and "smoking" also have a positive correlation equal to 0.65, showing that these two hazardous leisure activities often go hand in hand.

The matrix demonstrates that the symptoms correlate highly with some diseases. For instance, "chronic lung disease" correlates significantly with such symptoms as "coughing of blood" (0.71), "wheezing" (0.61), and "shortness of breath" (0.50). This means that such symptoms are more probable among people who suffer from chronic lung disease. On the other hand, there are health indicators that correlate rather weakly with some symptoms; such health indicators can be indicators of rather weak associations or the presence of other factors determining these symptoms. One of the important observations is the relationship between "genetic risk" and other variables. For instance, "genetic risk" is positively correlated with "obesity" at 0.81 and "chronic lung disease" at 0.68, indicating that a person who is at a higher genetic risk is more likely to possess obesity or chronic lung diseases. Lastly, "air pollution" is very much correlated with "genetic risk" at 0.75, indicating possible interaction between environmental exposure and genetic predisposition.

The matrix contains a variable called "level," which is assumed to be the global severity of a condition or disease stage. Several variables have very high positive correlations with "level," including "obesity" at 0.83, "air pollution" at 0.64, and "genetic risk" at 0.70. High positive correlations imply that subjects with higher exposure to these factors or who have those characteristics may be at higher risk of developing the health condition under study to a greater severity.

On the whole, the correlation matrix would yield extremely significant insight into how every factor from demographics and lifestyle habits to genetic risks and health symptoms are connected. Such associations might be useful in pinpointing risk factors and how different variables influence health. The strong correlation between variables like "obesity," "chronic lung disease," "smoking," and "level" points to a need to target such factors in disease prevention and management. The matrix will be a wonderful starting point for further research into causative relationships and target interventions toward its improvement for better health.

The matrix contains a variable called "level," which is assumed to be the global severity of a condition or disease stage. Several variables have very high positive correlations with "level," including "obesity" at 0.83, "air pollution" at 0.64, and "genetic risk" at 0.70. High positive correlations imply that subjects with higher exposure to these factors or who have those characteristics may be at higher risk of developing the health condition under study to a greater severity.

On the whole, the correlation matrix would yield extremely significant insight into how every factor from demographics and lifestyle habits to genetic risks and health symptoms are connected. Such associations might be useful in pinpointing risk factors and how different variables influence health. The strong correlation between variables like "obesity," "chronic lung disease," "smoking," and "level" points to a need to target such factors in disease prevention and management. The matrix will be a wonderful starting point for further research into causative relationships and target interventions toward its improvement for better health.

Figure 3: Frequency distribution of data for the top 10 attributes. The histogram of the variable "Age" shows a nearly normal distribution, peaking at 30-40 years. The sample mean of the age (μ) is approximately 37.2 years with a standard deviation of σ=12, meaning that the data falls roughly around this central value for age. The histogram of "Gender" looks very skewed to the right and possesses two substantially prominent values, which might be an enumeration of binary gender coding, like 1 for male and 2 for female, for example. The mean value of 1.4 implies there could be a gender imbalance in the dataset.

The histograms for "Air Pollution" and "Alcohol Use" reflect some very interesting features. "Air Pollution" has a mean of 3.8 with a standard deviation of 2.0; this indicates moderate exposure levels in the population with a slightly skewed distribution. "Alcohol Use" has a higher range with a mean at 4.6 and a standard deviation at 2.6, which showed a large variation in alcohol consumption patterns.

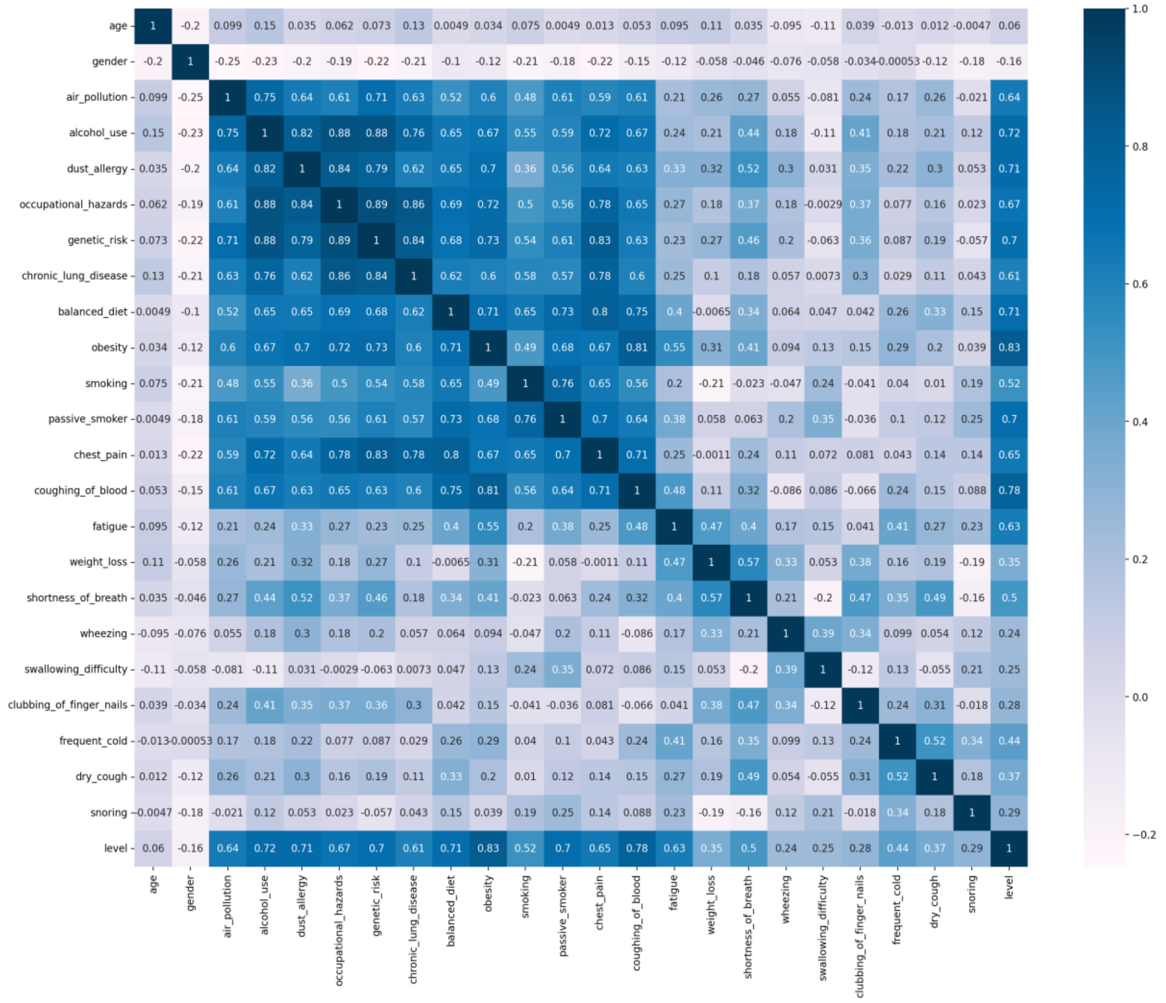

Figure 2: Correlation of all attributes

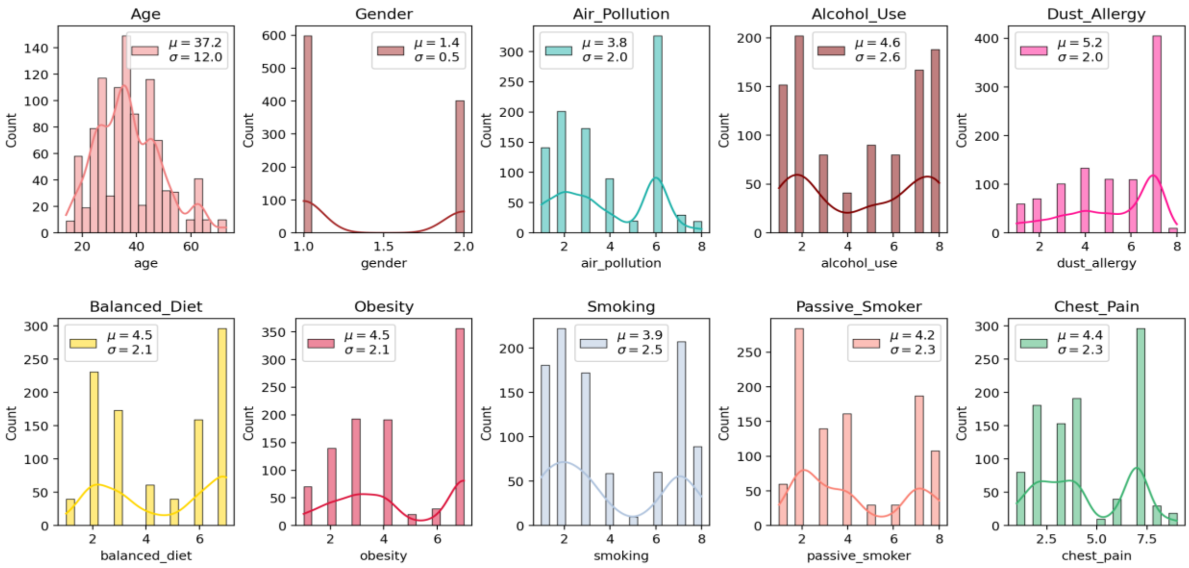

Figure 3: Histogram showing the contribution of attributes

TABLE 3: Summary of results obtained

Prediction Algorithm	Multinomial Logistic Regression			Random Forest Classifier			Naive Bayes			K-Nearest Neighbours (KNN)			Support Vector Machines(SVM)		
Metric \ Severity	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**
Precision	0.91	0.95	0.96	0.97	0.94	0.98	0.95	0.82	0.94	0.98	0.91	0.96	0.97	0.94	0.95
Recall	0.95	0.95	0.94	0.97	0.96	0.96	1.00	0.91	0.85	0.94	0.96	0.95	0.94	0.96	0.96
f1-Score	0.93	0.95	0.95	0.97	0.95	0.97	0.97	0.86	0.89	0.96	0.94	0.96	0.95	0.95	0.96
Accuracy	0.95			0.97			0.91			0.95			0.96		

The "Obesity," "Smoking," and "Chest Pain" distributions display interesting patterns. The "Obesity" distribution is somewhat balanced around a mean of 4.5, though spread considerably, providing a great indication that obesity levels generally varied within the population. The mean for "Smoking" is at 3.9 and the standard deviation at 2.5, which again allows one to understand that there is variation in smoking conditions. Likewise, "Chest Pain" presents multiple peaks-thus again indicating the presence of subgroups with varying levels of chest pain symptoms within the dataset.

Other histograms are "Passive Smoker" and "Dust Allergy". Based on these other histograms, we will be able to gather many different insights in terms of risk factors. For example "Passive Smoker" plots the exposure level with a mean value at 4.2 and a standard deviation of 2.3, hence passively smoking turns out to be a significant risk for a huge proportion of respondents in the data set. Similarly, "Dust Allergy" also presents some variation with a mean value at 5.2, showing just how common this is as an environmental health risk. Together, these visualizations would collectively provide a vivid description of the distribution patterns of the variables for demographics, environment, and health.

4. PERFORMANCE METRICS

Performance metrics are tools applied in quantification to determine the performance capabilities of predictive models or classification systems. Such metrics can indicate how good a model is at prediction by reviewing various aspects of its predictions against actual known outcomes.

Summary result for lung cancer severity classification with machine learning algorithms; Multinomial Logistic Regression, Random Forest Classifier, Naive Bayes, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM); are given in later sections in terms of precision, recall, f1-score, and accuracy classified into three degrees of severity: Low, Medium, High. The best classifier observed is the one with the random forest classifier at 0.97 Low, 0.94 Medium, and 0.98 High precision. Recall and f1-scores are always strong at all levels of severity. This means that the Random Forest model is very robust in establishing case classification across the different levels of severity with an all-time balanced performance concerning precision and recall.

The Multinomial Logistic Regression model also performed well, with precision ranging from 0.91 to 0.96 and recall consistently at or above 0.94. This model demonstrated strong accuracy in distinguishing between severity levels, making it a reliable choice, especially for applications where interpretability is crucial. The f1-scores for Logistic Regression remained high across all severity levels, indicating that it maintains a good balance between precision and recall, making it robust for predicting lung cancer severity.

Naive Bayes, KNN, and SVM also produced competitive results but with some variability. Naive Bayes, while achieving perfect recall (1.00) for Low severity, showed lower precision for Medium severity (0.82), which impacted its f1-score for that category. KNN and SVM both demonstrated high precision and recall, with KNN excelling in Medium severity recall (0.96) and SVM maintaining strong performance across all metrics. However, the lower recall for Naive Bayes in the High severity category and the slightly lower precision for KNN in Medium severity suggest that these models might be less consistent than Random Forest and Logistic Regression.

Overall, Random Forest emerged as the most reliable model, followed closely by Logistic Regression, with both providing high accuracy and balanced performance across all severity levels. The results obtained are summarized in Table 3.

5. CONCLUSION AND FUTURE WORK

In conclusion, the comparative analysis of various machine learning algorithms for lung cancer severity classification reveals significant insights into their respective performances. The Random Forest Classifier emerged as the standout model, demonstrating superior precision, recall, and f1-scores across all severity levels, making it the most reliable choice for this particular classification task. Multinomial Logistic Regression also showed impressive results, offering a strong balance between accuracy and interpretability, which could be particularly valuable in clinical settings where model explain ability is crucial. While Naive Bayes, K-Nearest Neighbors, and Support Vector Machines all produced competitive results, they showed some inconsistencies across different severity levels, potentially limiting their reliability in certain scenarios.

The five-year relative survival rate for lung cancer has seen gradual improvements, largely due to advancements in treatment options such as targeted therapies and immunotherapies. Despite these gains, lung cancer survival rates continue to trail behind those of many other cancers, highlighting the ongoing need for further research and innovation in this field.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

REFERENCES

[1]World Cancer Day 2024, Close the care gap- Addressing cancer care in India, 04th - February – 2024,

[2]American Cancer Society, Cancer Facts & Figures 2024.

[3]C. Anil Kumar, S. Harish, Prabha Ravi, Murthy SVN, B. P. Pradeep Kumar, V. Mohanavel, et al., "Lung Cancer Prediction from Text Datasets Using Machine Learning", Hindawi BioMed Research International, vol. 2022, pp. 10. https://doi.org/10.1155/2022/6254177

[4] Maurya, S.P., Sisodia, P.S., Mishra, R. et al. Performance of machine learning algorithms for lung cancer prediction: a comparative approach. Sci Rep 14, 18562 (2024). https://doi.org/10.1038/s41598-024-58345-8

[5] Venkatesh, S. P., Raamesh, L. (2023). Predicting Lung Cancer Survivability: A Machine Learning Ensemble Method on Seer Data. Int J Cancer Res Ther, 8(4), 148-154.

[6] Alam, Janee & Alam, Sabrina & Hossan, Alamgir. (2018). Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifie. 1-4. 10.1109/IC4ME2.2018.8465593.

[7]Raoof, Syed Saba, M. A. Jabbar, and Syed Aley Fathima. "Lung Cancer prediction using machine learning: A comprehensive approach." 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA). IEEE, 2020.

[8]M.V. Madhusudhan, V. Udayarani, and C. Hegde, "An intelligent deep learning LSTM-DM tool for finger vein recognition model USING DSAE classifier", Int. J. Syst. Assur. Eng. Manag., 2022.

[9]O. Hamzeh, A. Alkhateeb, J. Zheng, S. Kandalam, and L. Rueda, "Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data", BMC Bioinformat., vol. 21, no. S2, p. 78, 2020

[10]E. Shamsara, and J. Shamsara, "Bioinformatics analysis of the genes involved in the extension of prostate cancer to adjacent lymph nodes by supervised and unsupervised machine learning methods: The role of SPAG1 and PLEKHF2", Genomics, vol. 112, no. 6, pp. 3871-3882, 2020.

[11]K. L. Du, M. N. S. Swamy, "Particle Swarm Optimization", Search and Optimization by Metaheuristics,

Techniques and Algorithms Inspired by Nature, Springer International Publishing, 2016, pp. 153-173.

[12]Kaulgud, Sanjeev; Hulipalled, Vishwanath; Patil, Siddanagouda S.; Metipatil, Prabhuraj, "Detection of Prostate Cancer using Ensemble based Bi-directional Long Short Term Memory Network", Recent Advances in Electrical and Electronic Engineering, 2024, 17(1), pp. 91–98

[13]R. Kumar, P. Bhanti, A. Marwal, and R.K. Gaur, "Gene expression-based supervised classification models for discriminating early-and late-stage prostate cancer", Proc. Natl. Acad. Sci., India, Sect. B Biol. Sci., vol. 90, no. 3, pp. 541-565, 2020.

[14]T. Mehmood, A. Kanwal, M. M. Butt, “Naive Bayes combined with partial least squares for classification of high dimensional microarray data”, Chemometrics and Intelligent Laboratory Systems, Vol. 222, 2022, p. 104492.

[15]S. Begum, R. Sarkar, D. Chakraborty, S. Sen, U. Maulik, “Application of active learning in DNA microarray data for cancerous gene identification”, Expert Systems with Applications, Vol. 177, 2021, p. 114914.

[16]Kaulgud, S.P., Hulipalled, V., Patil, S.S., “Detection of prostate cancer related genes using modified ford-fulkerson algorithm in protein-to-protein interaction network”, International Journal of Engineering and Advanced Technology, 2019, 8(6), pp. 1595–1601

[17] S. M. D'Souza, S. H C, H. KG, N. Ashwin, V. KM and P. M. K S, "An Analysis of Smart Clinical Trial Investigation Methods in Research Medicine Industries," 2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP), Bengaluru, India, 2022, pp. 1-6.

[18] Metipatil, P., Bhuvaneshwari, P., Basha, S.M. et al. An Efficient Framework for Predicting Cancer Type Based on Microarray Gene Expressions Using CNN-BiLSTM Technique. SN COMPUT. SCI. 4, 381 (2023). https://doi.org/10.1007/s42979-023-01774-5