Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Early Detection and Monitoring of Respiratory Disorders using LASSO Regression on PPG Signals with Elephant Search Optimization

Manochithra A S.[1], Harikumar Rajaguru[2], Kalaiyarasi M.[3]

[1]PG Scholar, Department of CSE, Bannari Amman Institute of Technology, Sathyamangalam, India
[2]Professor, Department of ECE, Bannari Amman Institute of Technology, Sathyamangalam, India
[3]Associate Professor, Department of EIE, Bannari Amman Institute of Technology, Sathyamangalam, India

E-mail: [1]bala.manochitra@gmail.com, [2]harikumarr@bitsathy.ac.in, [3]kalaiyarasime@gmail.com

Abstract

Early diagnosis is the need of the hour in the treatment of respiratory-related health conditions. This study presents a novel method for monitoring respiratory disorders by applying a Least Absolute Shrinkage and Selection Operator (LASSO) regression model to Photoplethysmography (PPG) signals. By analyzing respiratory variations in the PPG waveform, the partial pressure of carbon dioxide (PCO_2) signal is extracted to monitor breathing patterns. The PCO_2 signal provides critical insights into respiratory dynamics, enabling the identification of irregular breathing rates and airflow obstructions. Using LASSO regression, the most relevant features from the PCO_2 signals are selected, reducing dimensionality and improving prediction accuracy. The proposed approach offers a cost-effective and non-invasive solution for evaluating respiratory health, making it suitable for both clinical and non-clinical settings. A comprehensive performance analysis demonstrates the efficacy of the LASSO regression-based method in diagnosing respiratory conditions. To evaluate its performance, five machine learning classifiers were employed: Linear Regression, Bayesian Linear Discriminant Analysis (BLDA), k-Nearest Neighbors (k-NN) with weighted voting, Expectation-Maximization (EM) with Logistic Regression, and Elephant Search Optimization (ESO). The results highlight the potential of this approach to improve healthcare by enabling early detection and management of respiratory disorders. The Elephant Search

DOI: https://doi.org/10.36548/jiip.2025.1.004
Received:07.02.2025, received in revised form: 03.03.2025, accepted: 17.03.2025, published: 28.03.2025

Optimization, combined with LASSO regression for dimensionality reduction, achieves 95.12% accuracy value, 95% F1 score, 0.90% MCC value, 4.87% error rate, 90.47% in Jaccard metrics, and 90% CSI.

Keywords: Elephant Search Optimization (ESO), LASSO, PCO_2, Bayesian LDA, PPG.

1. Introduction

Modern artificial intelligence (AI) and machine learning (ML) tools have significantly impacted the healthcare sector, particularly in the diagnosis of respiratory diseases. A systematic review conducted by Kapetanidis et al. [1] emphasizes the potential of AI-driven audio analysis, particularly through the examination of cough and lung sounds, to diagnose conditions such as MERS, COVID-19, asthma, and chronic obstructive pulmonary disease (COPD). The review consolidates findings from 75 research studies, revealing that models utilizing deep learning techniques including convolutional neural networks (CNN), and feature extraction methods like Mel-frequency cepstral coefficients (MFCC), have achieved high accuracy rates. Some of these models have demonstrated over 90% accuracy in classifying respiratory sounds. The COVID-19 pandemic has notably spurred research interest in this field, emphasizing the importance of standardized datasets and enhanced model generalizability to ensure dependable clinical implementation [1]. This system effectively identifies respiratory behaviors, such as normal breathing, coughing, and sneezing, by processing real-time data using techniques like wavelet and Fourier transforms. The classification is performed using a decision tree model, which improves recognition accuracy and highlights the system's capability to enable precise respiratory monitoring.

The sensor is cost-effective and can be easily integrated into everyday masks, making it highly suitable for widespread adoption in continuous health monitoring. Its non-invasive, real-time feedback is particularly beneficial in scenarios requiring consistent health tracking. Such advancements have the potential to significantly contribute to the early diagnosis and management of respiratory disorders, aligning with recent efforts to utilize machine learning for detecting respiratory conditions [2].This approach provides significant insights into diabetes detection by utilizing gene expression analysis for early identification. Through the application of dimensionality reduction and feature selection techniques, the study enhances diagnostic accuracy, equipping healthcare professionals with a robust tool to analyze extensive genetic datasets. In this framework, the optimized machine learning classifiers improve the

detection of type II diabetes, delivering a more accurate and dependable diagnostic solution [3].

This study assesses multiple machine learning classifiers for analyzing photoplethysmography (PPG) signals, which are widely used in cardiovascular monitoring. The research underscores the effectiveness of various algorithms in detecting cardiovascular diseases, with a particular focus on performance metrics including accuracy and sensitivity. The metrics are essential for enhancing the reliability and precision of diagnostic systems, ultimately leading to better cardiovascular disease detection as well as effective management [4]. The author investigates the effectiveness of different classifiers in detecting cardiovascular diseases (CVD) using photoplethysmography (PPG) signals. The study highlights that optimized classifiers significantly improve detection accuracy, emphasizing the potential of PPG as a non-invasive diagnostic tool for CVD. This research accentuates the importance of utilizing advanced machine learning models to enhance the diagnostic capabilities of PPG, making it a promising approach for early and non-invasive detection of cardiovascular conditions.

Respiratory disorders are of major concern, affecting millions of people and necessitating timely intervention for effective treatment, globally. The rise of non-invasive monitoring tools has underscored the importance of innovative approaches to assess respiratory health. Among these, machine learning-based processing of pCO_2 signals has gained attention due to its potential to enhance diagnostic accuracy. This study explores the integration of multiple machine learning classifiers, including Linear Regression, Bayesian Linear Discriminant Analysis (BLDA), k-Nearest Neighbors (k-NN) with weighted voting, and Expectation-Maximization (EM) with logistic regression, to monitor respiratory disorders using a LASSO regression model. Additionally, the study investigates the Elephant Search Optimization (ESO) classifier to optimize performance in detecting respiratory abnormalities. By using these advanced techniques, this research intends to contribute to the development of effective diagnostic tools that can significantly improve respiratory health monitoring, offering a non-invasive and precise approach to early detection and management of respiratory conditions.

2. Related Works

Li et al. [5] and colleagues conducted a significant study that integrated Convolutional Neural Networks (CNN) with Bidirectional Long Short-Term Memory (Bi-LSTM) techniques for the prediction and diagnosis of respiratory disorders. Their hybrid model achieved an accuracy of approximately 94%, demonstrating substantial improvements in diagnostic precision by effectively using both the temporal and spatial features of respiratory signals. These findings highlight the prospective combination of deep learning approaches for real-time disease monitoring and emphasize the value of such models in advancing healthcare applications. Hui et al. [6] and their team developed a machine learning algorithm based on LASSO regression to predict Post-Intensive Care Syndrome (PICS) associated with sepsis. Their algorithm achieved an accuracy of 88.5%, showcasing the effectiveness of LASSO regression in identifying critical features from complex datasets. This study emphasizes the importance of feature selection in enhancing model performance and offers valuable insights for clinical applications in critical care settings. By focusing on key predictors, the research highlights how machine learning can progress the early identification and management of PICS, eventually contributing to better patient outcomes in intensive care environments. Souvik Guha et al. [7] in his research focused on using LASSO regression for feature selection to boost the performance of deep learning simulations in diagnosing lung cancer from transcriptomic data. The study attained an impressive accuracy rate of 92%, demonstrating how optimizing feature selection can significantly improve the adeptness and effectiveness of diagnostic models. By identifying the most relevant features from complex transcriptomic datasets, Guha’s work emphasizes the important role of machine learning in advancing personalized medicine. This approach not only improves diagnostic precision but also paves the way for customized treatment strategies, highlighting the transformative potential of machine learning in therapeutic research and patient care. Shuzan et al. [8] and colleagues studied the application of machine learning for estimating respiration rates and blood oxygen saturation levels using photoplethysmogram (PPG) signals. Their research achieved an accuracy of 90.7% in estimating these vital signs, showcasing the potential of machine learning algorithms in non-invasive monitoring. The findings highlight the effectiveness of integrating machine learning techniques into PPG signal analysis, enabling more accurate and timely assessments of patients' respiratory health. This study highlights the promise of machine learning in enhancing non-invasive diagnostic tools, which can improve patient monitoring and

support early detection of respiratory abnormalities, ultimately contributing to better healthcare outcomes. In their research, Palanisamy and Rajaguru investigated the use of multiple classifiers to enhance the performance of detecting cardiovascular diseases (CVD) from photoplethysmogram (PPG) signals using machine learning methods. Their study achieved an overall accuracy of 87.5%, demonstrating the adaptability and effectiveness of machine learning in addressing complex healthcare challenges. By leveraging multiple classifiers, the researchers were able to improve the robustness and trustworthiness of CVD detection, highlighting the potential of machine learning in non-invasive diagnostic applications. The insights gained from this study contribute significantly to the ongoing development of intelligent healthcare solutions, particularly in the field of cardiovascular monitoring, leading to more accurate, efficient, and accessible medical devices[9].

3. Dataset Collection and Signal Preprocessing

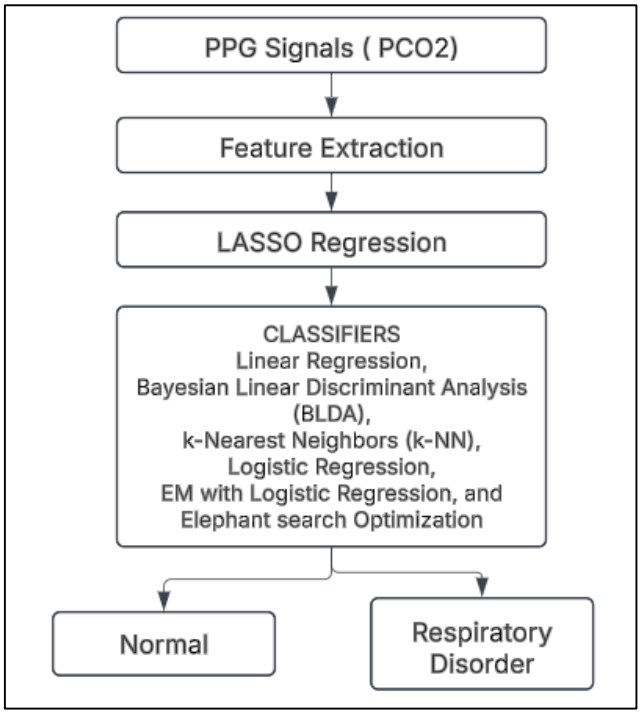

Figure 1. Schematic of the Methodology for Detection of Respiratory Disorder

Figure 1 indicates the schematic diagram of the methodology of the research with the application of LASSO as a dimensionality reduction method with mathematical and bio-inspired classifiers.

In this research, MATLAB 2017b was utilized for dataset preprocessing, feature extraction, and classification. The experiments were conducted on a desktop computer running a 64-bit Windows 10 operating system, equipped with an Intel(R) Core(TM) i3-3220 CPU @ 3.30 GHz and 8 GB of RAM.

3.1 Dataset Details

University of California datasets are used in the research. The PPG signals are re-sampled at 200Hz per second. Hence, there are 1, 44,000 per patient analyzed in this research. A collection of 41 subjects (20 with respiratory disease and 21 normal) was obtained from the Capno database. These PPG signals were labeled as normal and respiratory and were utilized for the evaluation of this work. The initial phase involves PPG signal acquisition. Providing dataset samples with feature details enhances clarity and improves reader understanding. Features such as signal amplitude, frequency components, and waveform characteristics were considered for analysis. The dataset consists of 41 subjects (20 with respiratory disease and 21 normal), labeled accordingly for evaluation.

3.2 Dimensionality Reduction by LASSO Regression

The study by Liu et al. [5] emphasizes the effectiveness of LASSO regression in extracting significant predictors from complex datasets, particularly in predicting postoperative lung complications in elderly patients. By reducing overfitting and focusing on the most relevant features, LASSO enhances the predictive accuracy of clinical models. This aligns with the \ application of LASSO regression to photoplethysmogram (PPG) signals for detecting respiratory disorders, demonstrating its versatility in analyzing physiological data. The research emphasizes the potential of LASSO with an intention to early detection and personalized risk assessment in healthcare, showcasing its value in addressing respiratory health challenges across diverse applications. These findings reinforce the importance of LASSO as a powerful tool for advancing precision medicine and improving patient outcomes. [10]

LASSO regression is highly effective for feature selection due to its use of regularization, which shrinks the coefficients of irrelevant variables to zero. This process not only simplifies the model but also enhances its accuracy, making LASSO particularly well-suited for handling high-dimensional datasets such as pCO_2 signals. By focusing on the most essential features, LASSO improves classifier performance and mitigates the risk of overfitting,

ensuring that the model generalizes well to new data. These properties make LASSO an invaluable tool for analyzing respiratory data, as it enables the identification of key predictors while maintaining model efficiency and interpretability. This approach is especially beneficial in healthcare appl ications, where accurate and reliable models are essential for early detection and diagnosis of respiratory conditions.

The LASSO objective function can be expressed mathematically as follows:

$$\beta = \frac{argmin}{\beta}\{\sum_{i=1}^{n}(x_i - \sum_{j=1}^{p}\beta_j y_{ij})^2\} + \gamma \sum_{j=1}^{p}|\beta_j| \quad (1)$$

where:

- y is the response variable,
- β is the intercept,
- $\beta_{,J}$ are the coefficients,
- x are the predictor variables,
- n is the number of observations,
- p is the number of predictors,
- γ is the regulation parameter that controls the strength of the penalty.

LASSO regression is a powerful tool for dimensionality reduction, particularly in scenarios where multicollinearity is present or where the number of predictors exceeds the number of observations. By leveraging its ability to perform variable selection and shrinkage, LASSO enables the creation of more interpretable and robust models. This makes it highly valuable across a wide range of applications, from healthcare, where it can improve diagnostic accuracy and biomarker identification, to finance, where it aids in risk modeling and portfolio optimization. Its ability to simplify complex datasets while maintaining predictive performance emphasizes its importance in modern data analysis.

After dimensionality reduction, the dataset size was reduced from 1,44,000 to 5,000 retaining the most relevant features while minimizing redundancy.

3.3 Statistical Analysis on PCO_2 Signals After LASSO Regression

In the field of respirational disorder detection, statistical analysis is essential for identifying and interpreting patterns within large patient datasets. Key statistical metrics, such as mean, variance, skewness, and kurtosis, are used to analyze the central tendency, variability,

and delivery of disease-related biomarkers or genetic factors. These metrics offer valuable insights into the characteristics of the data, enabling researchers to better understand the underlying patterns and trends associated with respiratory conditions. This analytical approach supports the development of more accurate diagnostic tools and personalized treatment strategies.

Pearson's correlation coefficient (PCC) is a vital tool for identifying relationships between variables, which can help uncover potential risk factors associated with respiratory disorders. Meanwhile, Canonical Correlation Analysis (CCA) is useful for understanding the relationships among multiple variables, offering deeper insights into the complex interactions underlying respiratory conditions. By applying these statistical methods and their associated formulas, researchers can develop predictive models that enhance the diagnosis, treatment, and overall understanding of respiratory diseases. These techniques play an important role in advancing respiratory health research and improving patient outcomes.

In addition, definitions and formulas for key statistical concepts, such as kurtosis, variance, skewness, mean, Pearson correlation coefficient (PCC), as well as canonical correlation analysis (CCA), are provided. These concepts are fundamental for analyzing data and building machine learning models, including LASSO regression, which is used for tracking and diagnosing respiratory disorders. By using these statistical tools, researchers can better understand data patterns, identify significant predictors, and develop robust models to improve the detection and management of respiratory conditions.

Mean (μ): The mean, commonly referred to as the average, is a numerical measure that epitomizes the central tendency of a dataset. The procedure requires summing all individual data values and subsequently dividing that sum by the total quantity of data points in the dataset.

$$\mu = \frac{1}{n}\sum_{i-1}^{n} x_i \tag{2}$$

Variance (σ^2): Variance enumerates the spread or dispersion of data points in a dataset. It represents the average of the squared deviations of each data point from the mean. This measure provides insight into the variability of the data.

$$\sigma^2 = \frac{1}{n}\sum_{i-1}^{n}(x_i - i) \tag{3}$$

Skewness (γ): Skewness is a numerical metric that describes the lopsidedness of a distribution around its mean. A positive skewness shows that the dispersal has an elongated or fatter tail on the right side, while a negative skewness suggests a longer or fatter tail on the left side. This measure helps in understanding the shape and symmetry of the data distribution.

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right) \tag{4}$$

Kurtosis (κ): Kurtosis is a statistical measure used to evaluate the shape of a distribution's tails in comparison to a normal distribution. It indicates whether the data has heavier tails (more outliers) or a sharper peak than a normal distribution. High kurtosis suggests a distribution with more extreme data points, while low kurtosis indicates a flatter distribution with fewer outliers. This metric helps in understanding the tail behavior and overall shape of the data.

$$K = \frac{\sum_{i=1}^{N}\frac{(x_i-\mu)^4}{N}}{\left(\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{N}\right)^2} - 3 \tag{5}$$

Pearson's Correlation Coefficient (PCC) is a statistical tool used to measure the strength and direction of a linear relationship between two variables, S and T. Its primary purpose is to compute the degree of linear association between the variables, indicating whether they move together (positive correlation), move in opposite directions (negative correlation), or have no linear relationship (zero correlation). The PCC range lies between -1 and 1. The values closer to 1 or -1 indicate stronger linear connections, while values near 0 suggest weak or no linear association.

$$p_{S,T} = \frac{\sum_{i=1}^{n}(s_i-s)(t_i-\mu\gamma)}{\sqrt{(\sum_{i=1}^{n}s_i-\mu_s)}} \tag{6}$$

Canonical Correlation Analysis (CCA) is a statistical method used to discover the relationship between two sets of variable quantity, X and Y. It recognizes linear combinations of variables within each set that capitalize on the correlation between the two groups. By solving a system of equivalences, CCA determines pairs of canonical variables that exhibit the highest correlation coefficient (r). The general form of the CCA equation for two datasets, X and Y, is designed to uncover the underlying relationships between the sets, providing insights

into their inter dependencies.This technique is particularly useful for analyzing complex multivariate data.

$$max_{a,b,c} corr(a^T X, b^T Y) \quad (7)$$

Table 1. Statistical Parameters for Lasso Regression Method

Statistical Parameters	**LASSO Regression**	
	Normal	**RD**
Mean	0.03727	-0.00414
Variance	0.00372	0.008551
Skewness	0.02332	0.084272
Kurtosis	0.02332	6.32719
Pearson Correlation Coefficient(PCC)	- 0.24946	-0.2566
Canonical Correlation Analysis (CCA)	0.79873	

Table 1 displays the statistical parameters for PPG signals after LASSO Regression as the Dimensionality Reduction method. It is identified from the above Table 1 that the Statistical parameters are very distinct across the two classes. The statistical parameters derived from LASSO regression highlight key differences between normal and respiratory disease (RD) PPG signals. The mean value for normal signals (0.03727) is higher than for RD signals (-0.00414), indicating an overall shift in signal distribution. RD signals exhibit higher variance (0.008551) compared to normal signals (0.00372), suggesting greater fluctuations in signal intensity. The skewness of RD signals (0.084272) is slightly higher than normal signals (0.02332), indicating more asymmetry in distribution. Additionally, RD signals have significantly higher kurtosis (6.32719) than normal signals (0.02332), implying the presence of more extreme values or peaks. The Pearson Correlation Coefficient (PCC) is negative for both normal (-0.24946) and RD (-0.2566) signals, with RD showing a slightly stronger negative correlation. Furthermore, the high Canonical Correlation Analysis (CCA) value (0.79873) suggests a strong relationship between the selected features and classification outcomes, demonstrating the effectiveness of LASSO in feature selection and dimensionality reduction. The negative PCC values show that the LASSO Regression values of the normal and Respiratory Disorder Classes are uncorrelated with loosely connected in nature. However, CCA values exhibits that there is a maximum correlation across the classes for LASSO Regression components of PCO_2 signals.

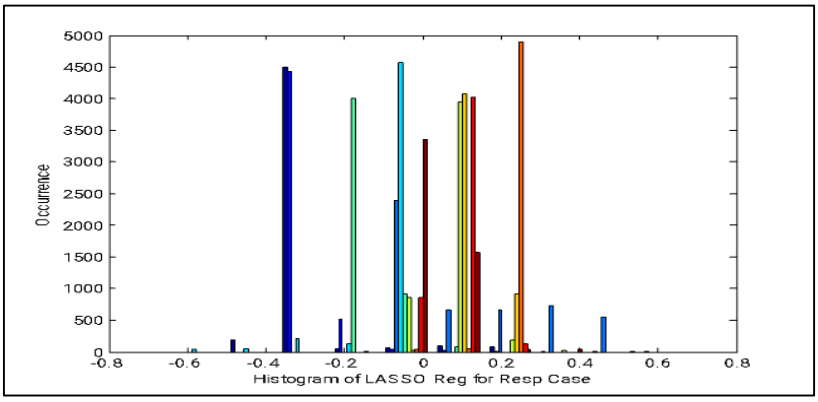

Figure 2. Histogram of Respiratory Cases in LASSO Regression

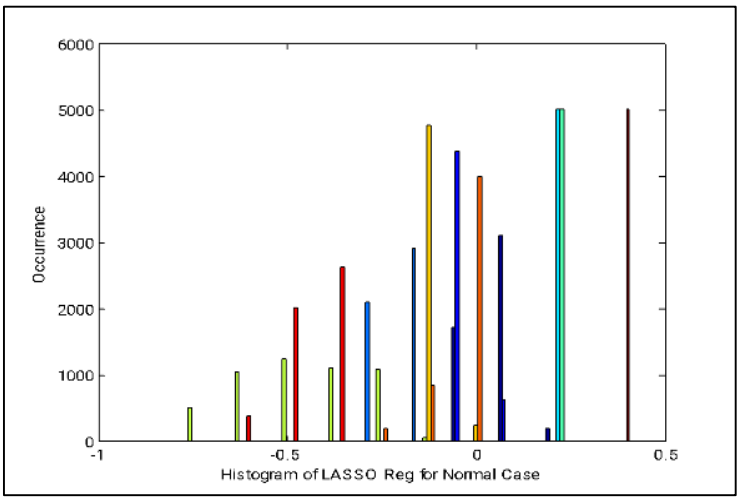

Figure 3. Histogram of Normal Cases in LASSO Regression

Figure 2 and Figure 3 depicts the histogram of LASSO Regression of the PCO_2 Waveforms. It is observed from the above figures that the histogram indicates that it is nonlinear , non-Gaussian, and with internal gaps with outliers variables. Even though, the CCA shows overlapping across the classes. Therefore, the selection of the classifier plays a vital role in the detection of respiratory disorder.

4. Role of Classifiers in Detection of Respiratory Disorders

In this section of the article discusses about the role of five machine learning classifiers in association with the detection of Respiratory Disorders from PPG signals.

4.1 Linear Regression

To predict postoperative lung function in patients with lung cancer, Kwon et al. [11] employed linear regression as a foundational model. This approach served as a benchmark to examine the relationship between clinical characteristics and lung function outcomes. While linear regression provided insights into straightforward feature-outcome relationships, it demonstrated limitations when applied to complex and nonlinear data. The study revealed that advanced models, such as Light GBM, outperformed linear regression in terms of accuracy and robustness. These findings underscored the need for more sophisticated methods in predictive tasks involving intricate datasets, highlighting the advantages of machine learning techniques over traditional statistical models in such scenarios.

Linear regression operates by minimizing the sum of squared differences between the observed and predicted values, identifying the line that best fits the data. The equation for linear regression is expressed as:

$$Y = X * \beta + \varepsilon \tag{8}$$

- Y denotes the predicted value.
- X is the input data (features).
- $\boldsymbol{\beta}$ represents the coefficients (weights) estimated from the data.
- ε is the error term, accounting for variance not explained by the model.

Linear regression is a commonly used technique across various fields for prediction and is frequently employed as a preprocessing step to explore relationships within data. By examining the coefficients and residuals, it provides insights into the relationships between predictors and the target variable. These insights can help refine more complex models or guide feature selection, ultimately enhancing the predictive accuracy of subsequent machine-learning applications. Its simplicity and interpretability make linear regression a valuable tool for initial data analysis and model development.

4.2 Bayesian Linear Discriminant Analysis (BLDA)

Linear Discriminant Analysis (LDA), also referred to as Normal Discriminant Analysis (NDA) or Discriminant Function Analysis (DFA), operates within a generative model framework. In this study, Bayesian Linear Discriminant Analysis (Bayesian LDA) is employed as a probabilistic classifier to identify lung conditions from chest X-ray images. By treating

class means and covariances as random variables with probability distributions, Bayesian LDA models uncertainty and provides probabilities for class membership.

The method emphasizes class separability by assuming equal priors for balanced datasets and adjusts priors based on class prevalence for unbalanced datasets to reduce bias. When applied to segmented images, Bayesian LDA improves prediction accuracy by focusing on disease-relevant features[12]. However, on unsegmented images, it handles broader features, which may introduce variability. The probabilistic outputs of Bayesian LDA support confident and interpretable clinical decision-making, making it a useful tool for analyzing medical imaging data.

The following is a mathematical representation of the linear discriminant function for two classes:

$$\delta(X) = X * (\sigma^2 * (\mu_0 - \mu_1) - 2 * \sigma^2 * (\mu_0^2 - \mu_1^2) + ln(P(w_0)/P(w_1))) \qquad (9)$$

where:

$\delta(x)$ represents the linear discriminant function.

x represents the input data point.

μ_0 and μ_1 are the means of the two classes. σ^2 is the common within-class variance.

$P(\omega_0)$ and $P(\omega_1)$ are the prior probabilities of the two classes.

4.3 K-Nearest Neighbors (k-NN) With Weighted Voting

The k-Nearest Neighbors (k-NN) method is a simple yet predominant machine learning algorithm used for both classification and regression tasks. As a non-parametric algorithm, it does not create an explicit model but as a substitute makes predictions based on the similarity of data points. An extension of the basic k-NN approach, known as weighted k-NN, will also be discussed.

Consider a dataset containing multiple data points, each associated with a specific class or value. When predicting the class or value of a new data point, k-NN identifies the k nearest neighbors (the most similar data points) in the training dataset. The prediction is then made based on the majority class in case for classification or the average value in case for regression, of these neighbors.

4.3.1 Weighted k-NN

The study "A Novel Feature-Significance Based k-Nearest Neighbour Classification Approach for Computer Aided Diagnosis of Lung Disorders" introduces adaptive weighted deviation-based metrics (AWDMs) to enhance the K-Nearest Neighbors (k-NN) algorithm for lung disease diagnosis. This approach assigns weights to each feature based on its significance in distinguishing between classes. A genetic algorithm is employed to optimize these weights, maximizing classification performance. By incorporating the weighted deviations into distance calculations, the method generates metrics such as Weighted City Block Distance (WCBD) and Weighted Euclidean Distance (WED). These weighted distances enable the k-NN algorithm to prioritize more informative features, significantly improving diagnostic accuracy. This innovative approach demonstrates the potential of feature-weighted k-NN in enhancing computer-aided diagnosis systems for lung disorders. [13]

$$P(Y = 1|X = \frac{1}{1+e^{-\beta 0+\beta 1+\beta 2+\cdots+}} \quad (10)$$

Where:

Y is the binary outcome variable,

X represents the predictor variables, and

β are the parameters to be estimated.

4.4 EM with Logistic Regression

In their study "Expectation Maximization Based Logistic Regression for Breast Cancer Classification," Rajaguru et al. [14] utilizes the Expectation Maximization (EM) method combined with Logistic Regression (LR) to enhance breast cancer classification. The EM algorithm iteratively refines the model's parameters to address missing or incomplete data. During the Expectation (E-step), missing data is estimated using the current model parameters. In the Maximization (M-step), the parameters are rationalized to maximize the likelihood of the observed data. By integrating EM with logistic regression, the model becomes more robust, improving classification accuracy and effectively handling incomplete datasets. This combination enables more accurate prediction of cancer types, particularly in noisy or ambiguous datasets, demonstrating its potential for advancing diagnostic precision in medical applications.

4.5 Elephant Search Optimization

The Elephant Search Optimization (ESO) is a bio-inspired optimization technique modeled after the behavior of elephants. It employs a population of search agents, referred to as elephants, to explore and exploit the search space. Figure 4 indicates the Flow chart for ESO as a Classifier. Initially, elephants scan the entire search space to identify potential solutions, and then they focus on refining these solutions in promising regions. This approach accomplishes a balance between global investigation and local manipulation, enabling the algorithm to efficiently locate optimal solutions for complex optimization problems.

ESO's adaptability to various problem settings allows it to outperform many traditional optimization methods, making it a powerful tool for solving challenging optimization tasks. [15].

$$X_i^{t+1} = X_i^t + \alpha . Direction. (X_{best} - X_i^{(t)}) \quad (11)$$

Where:

X is the position of elephant i at iteration t.

α is a step size parameter that controls the rate of movement.

X is the position of the best-performing elephant in the current iteration.

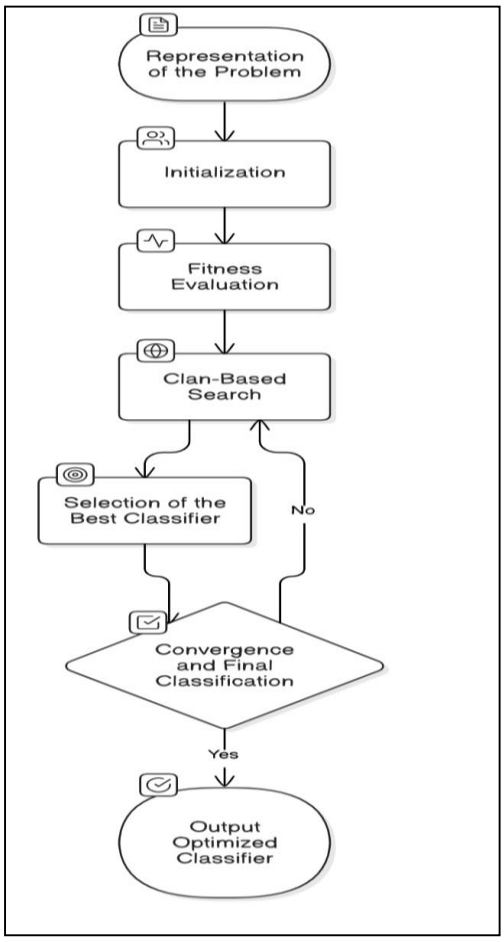

Figure 4. Flow Chart for Elephant Search Optimization as a Classifier

Direction in the ESO is represented as a vector that determines the movement of elephants within the solution space. This direction can be influenced by random values to promote exploration, ensuring a thorough search of the solution space. These mathematical principles form the foundation of the ESO, guiding the elephants' movements, evaluating candidate solutions, and enabling communication among the search agents. This framework allows the algorithm to efficiently optimize solutions in complex and high-dimensional problems, such as those involved in respiratory disorder monitoring. By balancing exploration and exploitation, ESO effectively navigates the search space to identify optimal solutions, making it a valuable tool for advanced classification tasks [16].

- **Initialize Population** – Generates multiple random feature subsets (elephants).
- **Evaluate Fitness** – Measures classification accuracy for each subset.
- **Global Search (Exploration)** – Moves elephants toward the best subset using

$$X_1^{new} = X + SFX\ (Xleader\ X_i) + a \times random$$

- **Local Search (Exploitation)** – Refines weaker subsets using:

$$X_1^{new} = X_{mean} + \beta X\ (X_{best} - X_i)$$

- **Feature Selection** – Keeps the most relevant features and remove redundant ones.
- **Repeat Until Convergence** – Continues iterations until accuracy stops improving.
- **Output Optimized Features** – The final reduced subset improves classification accuracy.

Exploration is performed by allowing elephant herds to search widely across the feature space, ensuring diverse candidate solutions and avoiding local optima. Exploitation refines the selected features by updating positions based on the strongest individuals, ensuring convergence to an optimal solution. This approach helps in reducing dimensionality while preserving the most relevant features, improving the efficiency and accuracy of respiratory disorder detection.

An initial dataset containing 144,000 features is reduced to 5,000 features after ESO-based selection. The classification accuracy improves from 85% to 92%, illustrating how ESO optimizes feature selection, reduces dimensionality, and enhances respiratory disorder classification accuracy.

4.6 Training and Testing of the Classifiers

As the number of patterns in each database for training is limited, the technique of S-fold cross-validation is employed to partition the data sets. The available data is split up into Subsets each of equal size. The first subset is chosen to be tested and the other S-1 subsets are combined to form the training and validation sets. After the network is trained using these, the classification performance of the test set is recorded. The process is then repeated so that each of the S-1subsets acts as the test set in turn. The final classification performance is the average of the S test set results. In this research, a value of fifteen was used for S per patient. This research associated with 41 patients each for normal and RD patients and multi trail training of classifiers is required. The use of cross-validation removes any dependence of choice of pattern for the test set. The training process is controlled by monitoring the Mean Square Error (MSE) which is defined as,

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(O_i - T_j)^2 \tag{12}$$

where Oi is the observed value at time i, Tj is the target value at model j; $j = 1$ and 2, and N is the total number of observations per epoch in the proposed research it is 5000. Based on the training and testing MSE values the classifier's parameters are optimized. The training process is controlled by max MSE of 10^{-05} or 400 iterations. Table 2 shows the training and testing MSE of the Classifiers.

Table 2. Average Training and Testing MSE of the Classifiers

Classifiers	**Average Training MSE**	**Average Testing MSE**
Linear Regression	7.29E-06	2.6E-05
Bayesian LDA	2.02E-06	2.12E-05
k-NN (Weighted)	8.41E-06	1.94E-05
EM with logistic Regression	5.76E-07	6.25E-06
Elephant Search Optimization	1.22E-08	4E-07

Table 2 depicts that ESO Classifier is place at lower training MSE of 1.22E-08 and average testing MSE of 4E-07. This implies that ESO Classifier may be settled at better classification accuracy than all other classifiers.

5. Result and Discussions

In monitoring respiratory disorders using pCO_2 signals, effective feature extraction is essential for the performance of machine learning classifiers such as LASSO regression. This study leverages various features directly derived from pCO_2 signals to capture patterns that differentiate between normal and respiratory disorder states.

To evaluate classifier performance, confusion matrices are employed. These matrices provide a summary of true and projected classifications, consisting of four key components: True Positives - TP, True Negatives - TN, False Positives - FP, and False Negatives - FN, allowing for a clear assessment of classification accuracy.

5.1 Performance Analysis of the Classifiers

The feature extraction process focuses on identifying relevant characteristics from both normal and respiratory disorder data. These extracted features are then fed into the LASSO regression model, along with other classifiers, to assess their performance.

Table 3. Performance Metrics

Performance Metrics	Derived from Confusion matrix
Accuracy	$ACC = \frac{TP + TN}{TP + FP + TN + FN}$
F1 Score	$F1 = \frac{2 * TP}{(2 * TP + FN + FP)}$
Mathews Correlation Coefficient	$MCC \frac{(TN * TP - FN * FP)}{\sqrt{((FP + TP) * (FP + TN) * (FN + TN))}}$
Error Rate	$ErR = \frac{(FN + FP)}{(TP + TN + FP + FN)}$
Jaccard Metric	$Jaccard = \frac{TP}{TP + FP + FN}$
CSI	$\mathbf{CSI = PPV + SEN - 100}$

Table 3 shows four key efficacy indicators used to evaluate the classifier's performance. These metrics demonstrate how effectively LASSO regression distinguishes between normal and respiratory disorder states, providing insights into the model's accuracy, precision, recall, and overall diagnostic capability. This approach ensures a robust evaluation of the classifier's ability to discriminate between healthy and pathological respiratory conditions.

Table 4. Performance Analysis of the Classifiers

Classifiers	Performance Metrics					
	Accuracy (%)	F1 Score (%)	MCC	Error Rate (%)	Jaccard Metric (%)	CSI (%)
Linear Regression	80.48	80.95	0.61	19.51	68.00	62.27
Bayesian LDA	85.36	85.71	0.71	14.63	75.00	71.81

k-NN (Weighted)	82.92	82.05	0.65	17.07	69.56	64.21
EM with Logistic Regression	87.80	87.17	0.75	12.19	77.27	74.47
Elephant Search Optimization	95.12	95	0.90	4.878	90.47	90

The Table 4 shows the evaluation performance of different classification algorithms for predicting respiratory conditions using PCO_2 signals. It uses metrics like Accuracy, F1 Score, Error Rate, Jaccard Index, and Critical Success Index (CSI) to show how well the algorithms classify respiratory patterns. These metrics give a clear picture of the models' strengths and weaknesses.

In the study of predicting respiratory disorders using PCO_2 signals, various classification algorithms were evaluated based on key performance metrics, including accuracy, F1 Score, Matthews Correlation Coefficient (MCC), Error Rate, Jaccard Metric, and Critical Success Index (CSI). Among these, the Elephant Search Optimization classifier exhibited superior performance, attaining an accuracy of 95.12%, an F1 Score of 95%, and a CSI of 90%. This highlights its strong ability to accurately distinguish respiratory patterns while maintaining a low misclassification rate (Error Rate: 4.878%). The Expectation-Maximization (EM) algorithm combined with Logistic Regression demonstrated notable performance, achieving an accuracy of 87.80%, an F1 Score of 87.17%, and a Critical Success Index (CSI) of 74.47%. These results suggest that it effectively balances precision and recall, making it a reliable choice for respiratory disorder classification. Similarly, the Bayesian Linear Discriminant Analysis (BLDA) attained an accuracy of 85.36%, an F1 Score of 85.71%, and a CSI of 71.81%, highlighting its effectiveness in accurately identifying respiratory conditions.

The Weighted k-Nearest Neighbors (k-NN) algorithm demonstrated solid performance, achieving an accuracy of 82.92% and a Critical Success Index (CSI) of 64.21%. However, it slightly trailed behind the Bayesian Linear Discriminant Analysis (BLDA), and EM with Logistic Regression in terms of effectiveness. On the other hand, Linear Regression, while still contributing valuable insights, produced comparatively lower results with an accuracy of 80.48% and a CSI of 62.27%, suggesting room for further enhancement. These results emphasize the Elephant Search optimization algorithm as the most effective choice for predicting respiratory disorders using PCO_2 signals. While other models also deliver

competitive performance, they may require additional refinement to optimize their effectiveness for specific applications. This analysis highlights the significance of selecting suitable classification algorithms to develop the accuracy and reliability of respiratory disorder monitoring in clinical settings.

6. Conclusion and Future Work

These research findings evaluated the performance of different classifiers in predicting respiratory disorders using PCO_2 signals. The study demonstrates that the Elephant Search Optimization (ESO) classifier is the most effective method for predicting respiratory disorders using PCO_2 signals, achieving the highest accuracy (95.12%), F1 Score (95%), and Critical Success Index (CSI) (90%). This highlights its superior capability in distinguishing respiratory patterns with minimal misclassification. Among them, the Elephant Search Optimization algorithm demonstrated the highest accuracy, with EM using Logistic Regression and Bayesian LDA also showing strong potential for healthcare applications. Future advancements will focus on refining these classifiers, investigating collaborative techniques, and exploring deep learning models to enhance accuracy and robustness. Furthermore, incorporating unconventional feature selection methods and using technical expertise could improve the reliability of respiratory disorder predictions in real-world clinical environments.

References

[1] Kapetanidis, Panagiotis, Fotios Kalioras, Constantinos Tsakonas, Pantelis Tzamalis, George Kontogiannis, Theodora Karamanidou, Thanos G. Stavropoulos, and Sotiris Nikoletseas. "Respiratory diseases diagnosis using audio analysis and artificial intelligence: a systematic review." Sensors 24, no. 4 (2024): 1173.

[2] Zhang, Chi, Lei Zhang, Yu Tian, Bo Bao, and Dachao Li. "A machine-learning-algorithm-assisted intelligent system for real-time wireless respiratory monitoring." Applied Sciences 13, no. 6 (2023): 3885.

[3] Chellappan, Dinesh, and Harikumar Rajaguru. "Detection of diabetes through microarray genes with enhancement of classifiers performance." Diagnostics 13, no. 16 (2023): 2654.

[4] Rajaguru, Harikumar, M. Gowri Shankar, S. P. Nanthakumar, and I. Arul Murugan. "Performance analysis of classifiers in detection of CVD using PPG signals." In AIP Conference Proceedings, vol. 2725, no. 1. AIP Publishing, 2023.

[5] Li, Li, Alimu Ayiguli, Qiyun Luan, Boyi Yang, Yilamujiang Subinuer, Hui Gong, Abudureherman Zulipikaer et al. "Prediction and Diagnosis of respiratory disease by combining convolutional neural network and bi-directional long short-term memory methods." Frontiers in public health 10 (2022): 881234.

[6] Hui, Kangping, Chengying Hong, Yihan Xiong, Jinquan Xia, Wei Huang, Andi Xia, Shunyao Xu, Yuting Chen, Zhongwei Zhang, and Huaisheng Chen. "LASSO-Based Machine Learning Algorithm for Prediction of PICS Associated with Sepsis." Infection and Drug Resistance (2024): 2701-2710

[7] Guha, Souvik. "Feature Selection Using Lasso Regression Enhances Deep Learning Model Performance For Diagnosis Of Lung Cancer from Transcriptomic Data." bioRxiv (2024): 2024-05.

[8] Shuzan, Md Nazmul Islam, Moajjem Hossain Chowdhury, Muhammad EH Chowdhury, Murugappan Murugappan, Enamul Hoque Bhuiyan, Mohamed Arslane Ayari, and Amith Khandakar. "Machine learning-based respiration rate and blood oxygen saturation estimation using photoplethysmogram signals." Bioengineering 10, no. 2 (2023): 167.

[9] Palanisamy, Sivamani, and Harikumar Rajaguru. "Machine learning techniques for the performance enhancement of multiple classifiers in the detection of cardiovascular disease from PPG signals." Bioengineering 10, no. 6 (2023): 678.

[10] Liu, Jie, Yilei Ma, Wanli Xie, Xia Li, Yanting Wang, Zhenzhen Xu, Yunxiao Bai, Ping Yin, and Qingping Wu. "Lasso-based machine learning algorithm for predicting postoperative lung complications in elderly: a single-center retrospective study from China." Clinical Interventions in Aging (2023): 597-606.

[11] Kwon, Oh Beom, Solji Han, Hwa Young Lee, Hye Seon Kang, Sung Kyoung Kim, Ju Sang Kim, Chan Kwon Park et al. "Prediction of postoperative lung function in lung cancer patients using machine learning models." Tuberculosis and Respiratory Diseases 86, no. 3 (2023): 203.

[12] Jasthy, Sreedevi, Krishnamurthy Ramasubramanian, Radhakrishna Vangipuram, Satyanarayana Bollu, and Krishnamurthy Ramasubramanian Sr. "Comparative Analysis of Machine-Learning Algorithms for Accurate Diagnosis of Lung Diseases Using Chest X-ray Images: A Study on Balanced and Unbalanced Data on Segmented and Unsegmented Images." Cureus 16, no. 1 (2024).

[13] Moon, K., and A. Jetawat. "Predicting Lung Cancer with K-Nearest Neighbors (KNN): A Computational Approach." Indian J. Sci. Technol 17, no. 21 (2024): 2199-2206.

[14] Rajaguru, Harikumar, and Sunil Kumar Prabhakar. "Expectation maximization based logistic regression for breast cancer classification." In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), vol. 1, IEEE, 2017. 603-606.

[15] Deb, Suash, Simon Fong, and Zhonghuan Tian. "Elephant search algorithm for optimization problems." In 2015 tenth international conference on digital information management (ICDIM), IEEE, 2015. 249-255.

[16] Singh, Harpreet, Birmohan Singh, and Manpreet Kaur. "An efficient feature selection method based on improved elephant herding optimization to classify high-dimensional biomedical data." Expert Systems 39, no. 8 (2022): e13038.