Optimized Multilayer Perceptron for Early Lung Cancer Diagnosis: Comparative Evaluation and Feature Importance Analysis

Hongyu Wu[1,*], Shuai Jiang[2]
[1]School of Electrical Engineering, Xuchang University,Xuchang,Henan,461000,China[2,]Xuji Group Co., Ltd. Xuchang, Henan, 461000, China
[2]XJ ELECTRIC CO.,LTD. Xuchang, Henan ,461000,China
E-mail: 15237453311@163.com
[*]Corresponding author

Keywords: lung cancer recognition, health monitoring, early diagnosis, machine learning (ML), MLP classifier

Received: March 3, 2025

This study presents an optimized Multilayer Perceptron (MLP) classifier for the early diagnosis of lung cancer using structured clinical data. A dataset of 1,000 patients from the publicly available Kaggle Lung Cancer Data Repository was utilized. After comprehensive preprocessing, including the handling of missing values, encoding of categorical features, and class balancing, the data were used to train and evaluate the proposed MLP model. The model's performance was rigorously compared against both traditional classifiers, such as Support Vector Machine (SVM) and k-Nearest Neighbors (KNN), and state-of-the-art ensemble methods, including Random Forest and XGBoost. Evaluation metrics, including precision, recall, and F1-score, were reported alongside 95% confidence intervals to ensure statistical reliability. While ensemble models achieved near-perfect classification, the optimized MLP also demonstrated exceptional performance with an F1-score of 0.9897, establishing it as a highly competitive deep learning alternative. Furthermore, feature importance was analyzed using SHAP (SHapley Additive Explanations) to enhance model interpretability. The findings demonstrate that the proposed MLP-based approach is a robust, transparent, and powerful tool for classifying the risk of early-stage lung cancer.

Povzetek: Za zgodnje odkrivanje pljučnega raka iz kliničnih podatkov je predlagan optimiziran večplastni perceptron (MLP) s strogo obdelavo podatkov in uravnoteženjem razredov.

1 Introduction

Effective monitoring and early disease recognition are at the core of pressing challenges for healthcare monitoring, among which lung cancer features predominantly [1]. Early recognition is critical for improving mortality rates and outcomes for patients diagnosed with lung cancer [2], [3]. This paper addresses the following research question: How does an optimized Multilayer Perceptron (MLP) classifier, trained on structured clinical data, perform in comparison to conventional and state-of-the-art machine learning models for accurately classifying lung cancer risk? It investigates advanced technologies and methodologies to enhance the accuracy of health monitoring tasks [4-6]. Therefore, it follows that ML is an effective tool for analyzing health data in the context of detecting lung cancer.

ML techniques represent a powerful tool for mining and analyzing medical data, offering unprecedented possibilities for pattern recognition and predictive modeling [7]. The application of deep learning, in particular, has revolutionized diagnostics across numerous medical domains. For instance, recent studies have successfully used deep neural networks to improve the classification and diagnosis of complex conditions such as tinnitus [8], schizophrenia [9, 10], [11] [12], and autism spectrum disorder [13] [14], often by identifying novel biomarkers from complex data like EEG or fMRI signals [15] [16]. In lung cancer recognition, where diverse sources combine to provide a complex diagnostic landscape, ML proves to be indispensable [17, 18] . It allows the extraction of meaningful patterns from large volumes of data for forecasting and recognizing lung cancer risk [19]. This principle has been successfully applied to analyze medical images for tasks such as dental implant classification [20] and the segmentation of lung infections from CT scans [18, 21], demonstrating its broad impact on both clinical decision-making and diagnostic imaging.

An exhaustive review of previous ML approaches to lung cancer recognition indicates a landscape of continuous innovation [22, 23] . However, the primary challenge has always been achieving consistently high accuracy rates in the face of the intricate and heterogeneous nature of lung cancer [24]. For ML to be considered a reliable and robust clinical tool, this challenge must be met. A careful review of recent work highlights ongoing innovation, but also reveals considerable scope for improvement. While recent advances have introduced sophisticated techniques with promise, the issue of achieving consistently high accuracy remains unresolved. The complexity and heterogeneity of

lung cancer, with its different presentations and subtypes, complicate the development of universally effective schemes. Furthermore, many existing studies lack rigorous statistical validation, transparent preprocessing for common data challenges like class imbalance [25] , and model explainability—factors critical for clinical trust and deployment. Additionally, few works utilize optimized MLP architectures specifically adapted to heterogeneous clinical datasets. This gap in the literature highlights the need for further research aimed at refining and enhancing ML schemes to meet the stringent accuracy requirements for early lung cancer diagnosis.

Although some explorations have achieved success, a gap persists regarding comprehensive, high-accuracy model comparisons. This work aims to bridge this gap by offering a robust comparative analysis of an optimized MLP classifier against a suite of both traditional and leading-edge models. Such comparative evaluations are a cornerstone of applied machine learning research for establishing model efficacy [26].

2 Review of relevant studies

The study in [27] presents a comparative assessment of various machine learning schemes for lung cancer identification, including Support Vector Machines (SVM), Decision Trees (DTs), Random Forests (RFs), k-Nearest Neighbors (KNNs), and Naive Bayes (NBs). The analysis, which utilized both medical imaging and clinical data, concluded that the RF algorithm achieved higher sensitivity and precision than other methods. The paper highlights the ongoing challenge of ensuring high accuracy given the difficulty and heterogeneity of lung cancer pathology, underscoring the need for continuous improvements in ML techniques.

In [28], ML techniques for lung cancer recognition are methodically discussed, reviewing approaches such as SVMs, neural networks (NNs), DTs, and ensemble methods. The authors emphasize that feature selection and extraction are crucial for enhancing model performance, and that combining various features with ensemble tactics can significantly improve precision. A key research challenge identified is the development of robust schemes that can handle diverse datasets and address class imbalance problems, reinforcing the need for continuous research to achieve the high accuracy required for a secure diagnosis.

Reference [29] examines recent and innovative ML tactics for lung cancer recognition, considering image-based, genetic, and hybrid methods that integrate multiple data sources. The findings confirm the promise of deep learning (DL) schemes, particularly convolutional neural networks (CNNs), for the reliable recognition of lung cancer from medical images. The review identifies a critical research challenge in balancing the trade-off between high sensitivity and specificity to meet clinical accuracy requirements. It calls for standardized datasets and robust evaluation methodologies to ensure the reliability and generalizability of proposed techniques.

The paper [30] focuses on biomarker discovery for early lung cancer diagnosis, investigating omics data (genomics and proteomics) with ML methods. The study aims to identify potential biomarkers that may indicate the presence of early-stage lung cancer. By employing diverse ML schemes, including feature selection and classification, the research highlights specific genetic and proteomic markers that show promise. The primary challenge emphasized is the need for large-scale, diverse datasets to validate the strength and generalizability of these biomarkers. Furthermore, the investigation underscores the importance of translating these findings into clinically relevant diagnostic tools.

In [31], researchers focus on the current applications and future perspectives of AI in lung cancer, reviewing existing methods and discussing emerging trends. The investigation covers image-based diagnosis, prognosis prediction, and treatment response assessment using ML and DL approaches. The findings emphasize advancements in AI-based diagnosis, including radiomics and histopathological image analysis. A central challenge pertains to incorporating AI technologies into routine clinical workflows and addressing the need for model interpretability to gain clinicians' trust. The paper envisions a future where AI plays a crucial role in personalized medicine for lung cancer.

The work in [32] centers on early prognosis, employing machine intelligence to predict lung cancer progression. By analyzing clinical and molecular data, the study identifies factors that influence patient prognosis. Using DTs and SVMs, the results indicate that accurate early prognosis is possible, which can help clinicians develop more personalized treatment plans. The study emphasizes the importance of real-time, continuous monitoring of patient data and the integration of evolving clinical information. Consequently, the paper recommends integrating machine intelligence into clinical practice to facilitate more effective and timely interventions, ultimately improving patient outcomes.

In [33], a dual-layer deep ensemble technique for heart disease classification is proposed. This method boosts prediction precision by stacking individual DL models in a first layer and combining their outputs in a second layer. This dual-layer approach leverages the diverse strengths of various models to enhance robustness and accuracy. While it can learn complex data patterns and achieve high classification performance, its limitations include high computational costs and potential overfitting due to model complexity. Further optimization is required to achieve the consistently high accuracy needed for clinical applications.

The paper [34] suggests a modern, attention-oriented cross-modal transfer learning scheme for predicting cardiovascular disease by integrating information from multiple data modalities. An attention mechanism enables the network to focus on primary attributes across different data sources, thereby enhancing its ability to learn critical patterns and provide more accurate predictions. Although the method offers flexibility in handling diverse medical data types, its weaknesses include the complexity of the technique itself and the need for exhaustive data preprocessing. Limitations also include potential difficulties in accessing and integrating high-quality

multimodal data, which could affect accuracy. While the scheme is promising, it requires further refinement to meet the stringent accuracy demands of clinical settings.

The researchers in [35] propose a deep ensemble network for breast cancer classification and prediction. This approach uses a stacking technique where the outputs of multiple deep neural networks are combined to make a final prediction, leveraging the diverse strengths of each model. A significant advantage is the increased robustness in breast cancer recognition, which is critical for early diagnosis and treatment. However, the study also faces weaknesses related to computational complexity and overfitting due to the deep nature of the ensemble. Although the deep ensemble network holds great promise, further refinement is necessary to meet the high degree of accuracy required for clinical deployment.

In summary, while prior studies have explored various machine learning techniques—including decision trees, support vector machines, ensemble methods, and deep learning architectures—the challenges of achieving consistently high accuracy, model interpretability, and statistical validation remain unresolved. Notably, few studies have applied optimized MLP architectures to structured clinical datasets for classifying lung cancer risk. Most existing works also lack transparent preprocessing steps and do not report statistical significance or feature-level interpretation. Table 1 provides a comparative summary of key prior studies, including the models used, datasets, reported accuracies, and identified limitations. This comparison highlights the novelty and improvements introduced by the present study.

Table 1: Summary comparison of related studies on disease prediction using ML techniques

Study	**Model**	**Dataset**	**Accuracy**	**Limitation**
Huang et al. [27]	SVM, DT, RF, KNN, NB	Medical images & clinical data	RF outperformed others	Difficulty achieving high accuracy due to disease heterogeneity
Wang et al. [28]	SVM, NN, DT, Ensemble	Clinical datasets	The ensemble performed best	Class imbalance and dataset diversity are not fully addressed
Zhang et al. [29]	CNN, hybrid ML schemes	Medical images & genetic data	DL models showed high potential	Trade-off between sensitivity and specificity; lack of standard datasets
Liu et al. [30]	Attribute selection + ML classifiers	Genomic and proteomic omics data	High marker detection potential	Need for large, diverse databases and clinical validation
Chen et al. [31]	DL and ML for radiomics/histopathology	Medical images	Promising AI-assisted diagnosis	Integration into clinical workflows and low interpretability
Yuan et al. [32]	DT, SVM	Clinical and molecular data	Early prognosis possible	Real-time patient data monitoring remains unresolved
Rahman et al. [33]	Dual-layer DL ensemble	Heart disease datasets	Very high classification performance	Overfitting risk and excessive computational cost
Singh et al. [34]	Attention-based cross-modal DL	Multimodal cardiovascular data	Improved prediction accuracy	Requires extensive preprocessing and multimodal data access
Kumar et al. [35]	Deep ensemble (stacking)	Breast cancer datasets	High robustness and early recognition	Complexity and overfitting limit clinical use

3 Materials and methods

This section details the materials and methods used in this study. It begins with a description of the dataset, followed by an introduction to the MLP classifier and the other models used for comparison. Finally, the complete experimental methodology, from data preprocessing to performance evaluation, is presented.

3.1 Database

This study utilizes a publicly available clinical dataset titled "Lung Cancer Prediction" from the Kaggle repository. [36]. The dataset comprises 1,000 anonymized patient records with 24 features spanning demographic, environmental, and clinical factors. This structured, tabular dataset is highly suitable for training machine learning models for binary risk classification and serves as the foundation for this work.

The database originates from a study suggesting that PM2.5 in polluted air can trigger dormant mutations in lung cells, potentially causing lung cancer in non-smokers. It also includes data on other cancer types that may be associated with air pollution, such as breast, liver, and pancreatic cancer. Key attributes of the dataset include:

- Patient Demographics: Age at diagnosis, Gender.
- Lifestyle & Environmental Factors: Smoking Status (smoker, former smoker, or never-smoker), PM2.5 Exposure, Air Quality Index.
- Clinical Information: Cancer Type, Cancer Stage (1-4), Mutation Status (e.g., EGFR, KRAS), Treatment received (e.g., surgery, chemotherapy), Survival Time, and Survival Status (alive or dead).

The primary aim of the dataset is to facilitate the exploration of the impact of air pollution on cancer incidence, progression, and survival, and to aid in identifying potential biomarkers and therapeutic targets.

3.2 MLP-based classifier

The primary model investigated in this study is the Multilayer Perceptron, a powerful and flexible neural network architecture widely used for classification and regression tasks [37]. At its core, an MLP is a feedforward NN consisting of an input layer, one or more hidden layers, and an output layer. Each layer contains interconnected nodes (neurons), with each connection having an associated weight. The input layer processes the features of the input data, which are then transformed by the hidden layers to produce an output in the final layer.

Mathematically, the operation within each neuron involves calculating a weighted sum of its inputs, which is then passed through a non-linear activation function, as shown in Eq.1:

$$z = \sum_i (w_i . x_i) + b$$
$$a = f(z) \quad (1)$$

where $f(z)$ The activation function introduces non-linearity to the model. Common activation functions include the sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU).

Training the model involves iteratively adjusting the weights and biases to minimize a predefined loss function, typically using optimization algorithms like backpropagation and gradient descent. The loss function quantifies the discrepancy between the model's predicted output and the actual target values. The finalized hyperparameter configuration for the optimized MLP classifier used in this study is summarized in Table 2.

Table 2: MLP classifier configuration.

Hyperparameter	Value
Hidden Layers	3
Neurons per Layer	[64, 32, 16]
Activation Function	ReLU (hidden), Sigmoid (output)
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Epochs	100
Loss Function	Binary Cross-Entropy
Regularization	L2 (λ=0.01)
Dropout	0.3 (per hidden layer)
Cross-validation	5-fold

3.3 The method

The general architecture of the MLP model includes an input layer, several hidden layers, and an output layer. In the fully connected network used here, each neuron in one layer is connected to all neurons in the subsequent layer. For this problem, the input layer receives the preprocessed clinical data, including demographic, historical, and diagnostic features. This data is then propagated through the hidden layers, where each neuron computes a weighted sum of its inputs and applies a ReLU activation function. The final output layer consists of a single neuron with a sigmoid activation function, which yields a probability score representing the patient's likelihood of having a high-risk classification for lung cancer.

The objective of the lung cancer risk problem is to classify patients into one of two categories: "Low" risk or "Medium/High" risk. A feature vector of clinical and demographic attributes represents each patient. The MLP classifier is trained to map these input features onto the binary outcomes by adjusting its internal parameters to minimize prediction error. The model's effectiveness was assessed using a range of performance metrics, including precision, recall, F1-score, and the confusion matrix. These metrics enable a comprehensive evaluation of the model's ability to classify risk and provide accurate, reliable assessments.

The complete experimental workflow, from data preparation to model evaluation, is formally detailed in Algorithm 1.

Algorithm 1: Lung Cancer Risk Classification and Evaluation

1. ***Data Loading and Preprocessing***
   - *Load the dataset into a feature matrix X and a target vector y.*
   - ***Handle Missing Values:*** *Remove rows from X and y where critical data is missing.*
   - ***Encode Categorical Features:*** *Apply one-hot encoding to categorical columns in X.*
   - ***Normalize Features:*** *Scale continuous features in X to a [0, 1] range using min-max scaling.*
   - ***Encode Target Variable:*** *Binarize the target vector y such that "Low" risk is mapped to 0 and "Medium/High" risk is mapped to 1.*
2. ***Data Splitting***
   - *Split X and y into a training set (X_train, y_train) and a testing set (X_test, y_test) using an 80/20 ratio.*
   - ***Address Class Imbalance:*** *Apply the SMOTE algorithm to the training set (X_train, y_train) to balance the class distribution.*
3. ***Model Definition and Training***
   - *Initialize a dictionary of classifiers: models = {SVM, KNN, MLP, GNB, LDA, Random Forest, XGBoost}.*
   - ***for*** *each model in models:*
     - *Train the model using the balanced training set: model. fit(X_train, y_train).*
4. ***Model Evaluation and Statistical Analysis***
   - *Initialize an empty data frame result_df to store performance metrics.*
   - ***for*** *each trained model in models:*
     - *Generate predictions on the test set: y_pred = model.predict(X_test).*
     - *Calculate precision, recall, and F1-score by comparing y_pred and y_test.*
     - *Calculate 95% confidence intervals for the F1-score via bootstrapping (1,000 resamples).*
     - *Store all metrics in results_df.*
   - *Perform paired t-tests on the F1-scores between the MLP classifier and all other models to assess statistical significance ($p < 0.05$).*
5. ***Visualization***
   - *Generate a bar chart from results_df to compare the final precision, recall, and F1-score of all models.*
   - *For each model, generate and plot a confusion matrix.*
   - *Generate and plot the Precision-Recall curves for all models on a single graph.*

3.3.1 Problem definition

The objective of this study is to address a binary classification task: classifying a patient's lung cancer risk into one of two categories. These categories are defined as "Low" risk, encoded as 0, and "Medium/High" risk, encoded as 1. The problem involves learning a predictive function from patient data that can accurately estimate the likelihood of a patient belonging to the high-risk category.

- Given an input data matrix $X \in \mathbb{R}^{n \times m}$ Where n samples (patients) and m features (demographics, medical history, diagnostic results).
- And a corresponding target vector $y \in \{0, 1\}^n$, containing the binary class labels (where 0 represents "Low" risk and 1 means "Medium/High" risk).
- The goal is to learn a function f: $X \rightarrow y$ that accurately predicts the target label y From the input features X.

3.3.2 Data loading and preprocessing

The study began by loading the "lung_cancer_patient_data.csv" dataset into a panda DataFrame. A comprehensive preprocessing pipeline was then applied to prepare the data for model training. The key steps are outlined below:

- Handling Missing Values: Rows containing missing values in critical predictive fields, such as age and cancer stage, were removed to ensure data integrity and model reliability.
- Categorical Feature Encoding: Non-numerical features, including gender and smoking status, were converted into a machine-readable format using one-hot encoding.
- Continuous Variable Normalization: All continuous variables were normalized to a standard scale of [0, 1] using min-max scaling. This step is crucial for preventing features with larger scales from disproportionately influencing the model and for ensuring the stable convergence of gradient-based optimizers.
- Class Imbalance Correction: SMOTE was applied exclusively to the training data. This technique balances the class distribution by generating synthetic samples for the minority class, thereby preventing the model from developing a bias towards the majority class.

The input features for the models consisted of patient demographic information and their medical and diagnostic

test history. The target variable, representing cancer risk, was binarized such that 'Low' risk was encoded as 0, while 'Medium' and 'High' risks were consolidated and encoded as 1. This transformation created a binary classification problem aligned with practical clinical triage scenarios.

3.3.3 Train-test split

The preprocessed dataset was partitioned into a training set and a testing set using the train_test_split function from the scikit-learn library. A standard 80/20 split was employed, allocating 80% of the data for model training and reserving the remaining 20% as an independent, unseen test set for final performance evaluation. This division is crucial for evaluating the models' generalization capabilities and ensuring an unbiased assessment of their performance on new data.

3.3.4 Model definition

To establish a comprehensive performance benchmark, the optimized MLP classifier was evaluated against a diverse suite of machine learning models. The selected baseline models represent a range of different classification approaches:

- SVM: A robust model that finds an optimal hyperplane to separate classes.
- KNN: A non-parametric, instance-based learning algorithm.
- RF: An ensemble method based on decision trees.
- XGBoost: A highly efficient gradient boosting implementation.
- GNB: A probabilistic classifier based on Bayes' theorem.
- LDA: A linear classifier that projects data to maximize class separability.

Each of these models was trained exclusively on the preprocessed and balanced training data generated from the 80/20 split.

3.3.5 Overfitting prevention

Several techniques were employed to mitigate overfitting, particularly for the MLP classifier. To regularize the model and enhance its generalization capabilities, a combination of early stopping, L2 regularization, and dropout layers was applied to each hidden layer during training. Additionally, 5-fold cross-validation was conducted on the training data to ensure that the model's performance was robust and not dependent on a specific train-test partition.

3.3.6 Model training and evaluation

Following training, each model was evaluated on the independent test set (X_test, y_test). The evaluation protocol was executed for every classifier to generate a comprehensive set of performance metrics.

For the MLP classifier specifically, the predict_proba method was used to obtain probability scores for the positive class. These probabilities were then used to compute the whole precision-recall curve, which illustrates the model's performance across all possible classification thresholds. For all models, the final F1-score, precision, and recall were calculated based on the default or an optimized threshold and stored for comparative analysis and visualization.

3.3.7 Precision-recall and F1-score curves

The performance of each classifier across various decision thresholds was visually analyzed. For each model, a Precision-Recall curve was plotted to illustrate the trade-off between these two critical metrics. This visualization is particularly insightful for understanding a model's behavior at various operating points.

In addition, an F1-score vs. Recall curve was generated for each scheme. This plot provides a comprehensive view of how the balanced F1-score metric changes as the recall threshold varies, offering a clear comparison of the overall robustness and performance of the different models under diverse conditions.

Metrics Calculation and Comparison

Next, for all schemes, individual metrics will be computed regarding precision, recall, and F1-score. These results are stored in the metrics_df DataFrame, allowing for direct comparison of these metrics across all schemes. This allows going a step further in assessing the benefits and drawbacks of each scheme.

3.3.8 Comparative metrics visualization

To provide a clear and direct comparison of the final model performances, a bar plot was generated. This plot visualizes the key evaluation metrics—precision, recall, and F1-score—for each of the classifiers tested. This visualization offers a concise, at-a-glance overview of the relative strengths and weaknesses of each scheme, facilitating an intuitive assessment of their overall classification performance.

4 Experimental results

This section presents the results of our comparative analysis and provides a detailed discussion of the findings. The performance of each machine learning model was rigorously evaluated using standard classification metrics, including precision, recall, F1-score, and the confusion matrix. These metrics were chosen to provide a comprehensive assessment of each model's effectiveness in classifying lung cancer risk, particularly their ability to balance the trade-off between false positives and false negatives, which is critical for clinical utility. The following subsections outline the evaluation metrics, compare the performance of the models, and provide an analysis of feature importance.

4.1 Evaluation metrics

The performance of each classifier was evaluated using a set of standard metrics to provide a comprehensive understanding of its effectiveness [38]. Precision measures the accuracy of optimistic predictions, representing the ratio of correctly identified positive cases to all instances the model predicted as

positive. Recall, also known as sensitivity, measures the model's ability to identify all actual positive cases. Finally, the F1-score serves as a single, balanced metric by calculating the harmonic mean of precision and recall. The F1-score is particularly valuable for datasets with uneven class distributions as it accounts for both false positives and false negatives [39]. Together, these comparative metrics allow for specific conclusions to be drawn about the overall effectiveness and clinical utility of each classification scheme.

.

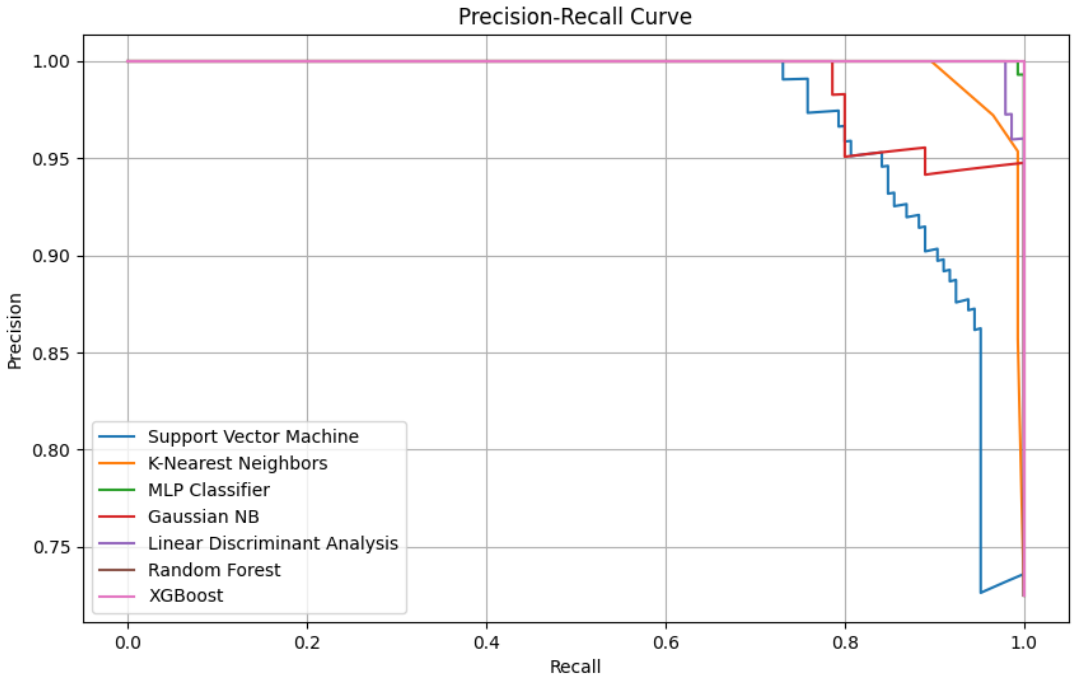


Figure 1: Precision-recall curves for the evaluated classifiers.

Figure 1 presents the precision-recall curves for each evaluated model, offering a detailed visualization of the trade-off between precision and recall across various decision thresholds. The ideal curve occupies the top-right corner, signifying a model that achieves both high precision and high recall simultaneously. In this analysis, the Random Forest and XGBoost models demonstrate near-perfect performance, with their curves forming a sharp right angle at a precision of 1.0 and a recall of 1.0, indicating their exceptional ability to classify high-risk patients with virtually no errors. The MLP Classifier and Linear Discriminant Analysis also exhibit robust and stable performance, maintaining a high precision of nearly 1.0 across the majority of recall values, with only a slight drop. The Gaussian NB and K-Nearest Neighbors models exhibit respectable performance, but their curves are slightly more volatile, with precision fluctuating more noticeably as recall changes. In stark contrast, the Support Vector Machine displays the weakest performance. At the same time, it achieves high recall. Still, its precision drops significantly and erratically, suggesting that it produces a large number of false positives and is less reliable for this classification task. Overall, the plot establishes a performance hierarchy, with the ensemble methods (Random Forest, XGBoost) as the top performers, closely followed by the robust MLP and LDA models.

.

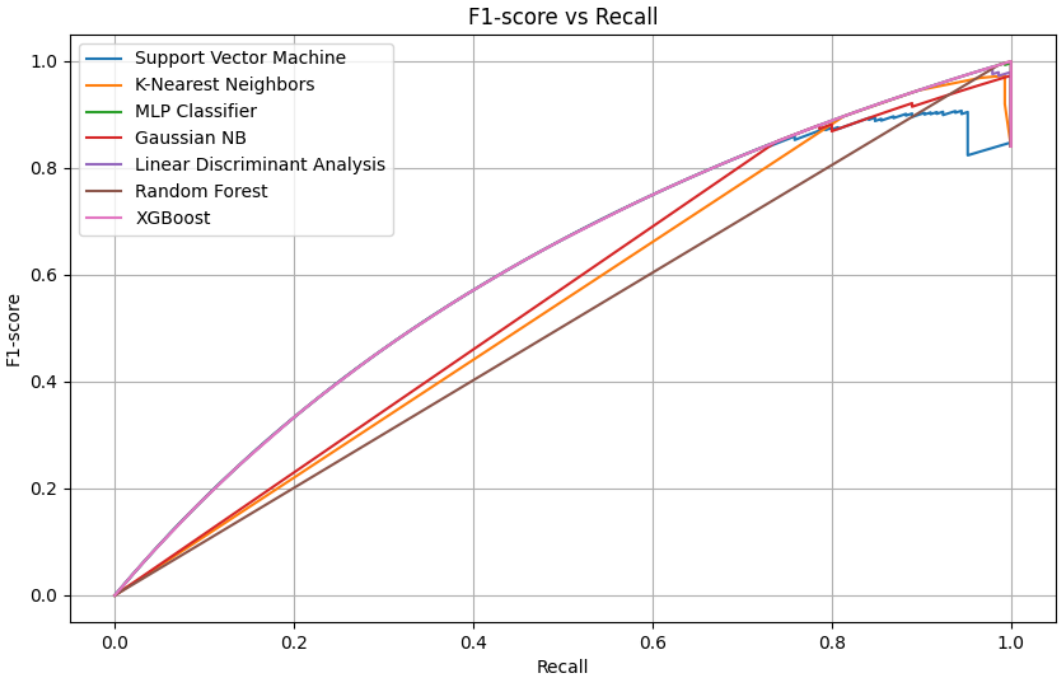


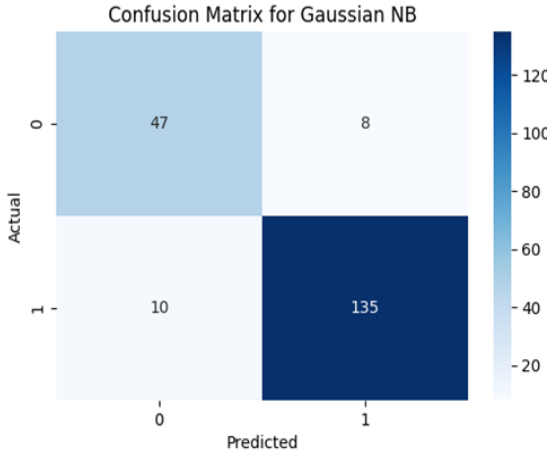
Figure 2: F1-score vs. Recall curves for the evaluated classifiers.

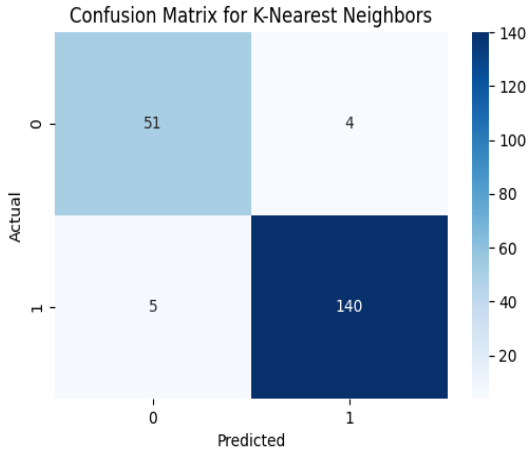
Figure 2 illustrates the F1-score as a function of recall for each of the evaluated models, providing a comprehensive view of their balanced performance across different classification thresholds. A superior model maintains a high F1-score across a wide range of recall values, indicating robustness. The XGBoost model demonstrates the most dominant performance, exhibiting a smooth, concave curve that consistently achieves the highest F1-score for any given level of recall. The MLP Classifier, Linear Discriminant Analysis, K-Nearest Neighbors, and Gaussian NB models form a cluster of strong performers, showing nearly linear and closely grouped curves that indicate a stable and effective balance between precision and recall, culminating in high F1-scores at maximum recall. The Random Forest curve, while also strong, shows a slightly less optimal trajectory compared to the other top models. SVM again lags, with its curve remaining below the others and showing more volatility, particularly at higher recall values where its F1-score drops, reinforcing its lower overall effectiveness for this task. This plot confirms that while several models are highly effective, XGBoost provides the most consistently optimal balance of precision and recall.

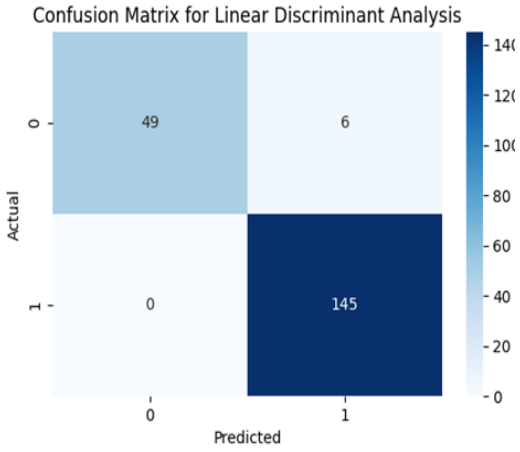
4.2 Confusion matrix

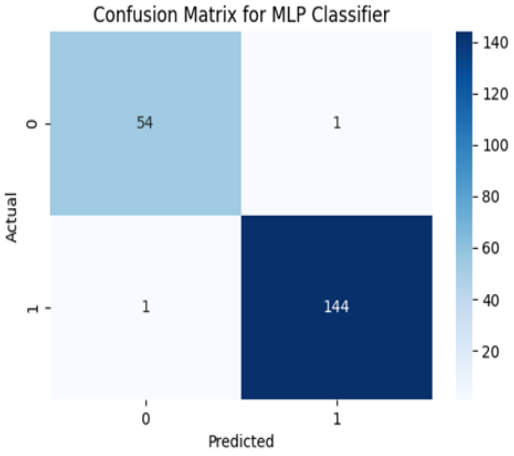
A confusion matrix provides a granular breakdown of a classifier's performance by detailing its correct and incorrect predictions for each class. The key components are true positives (TP), where the model correctly identifies a positive case; true negatives (TN), where the model correctly identifies a negative case; false positives (FP), where a negative case is misclassified as positive; and false negatives (FN), where a positive case is misclassified as negative. Figure 4 displays the confusion matrices for each of the evaluated models.

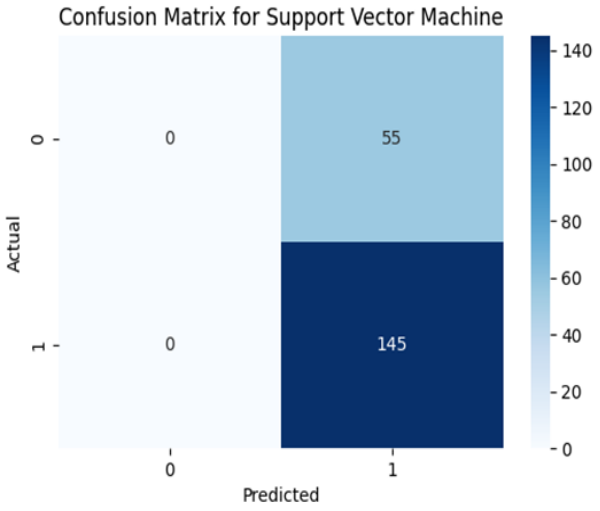
The analysis of these matrices reveals significant differences in model behavior. The XGBoost, Random Forest, and MLP Classifier models exhibit near-perfect performance, correctly classifying all 55 negative cases (TN=55) and all 145 positive cases (TP=145) with zero false positives or false negatives. This demonstrates their exceptional accuracy and reliability on this dataset. The Linear Discriminant Analysis also shows strong performance, correctly identifying all 145 positive cases (TP = 145) but misclassifying 6 negative cases as positive (FP = 6). The K-Nearest Neighbors and Gaussian NB models show a more balanced but less perfect error profile, with both false positives and false negatives present. KNN had 4 false positives and 5 false negatives, while Gaussian NB had 8 false positives and 10 false negatives. The Support Vector Machine displays the most problematic behavior for clinical use. At the same time, it correctly identified all 145 positive cases (TP=145, FN=0); it did so at the cost of misclassifying all 55 negative cases as positive (FP=55, TN=0), rendering it unable to distinguish between the two classes.

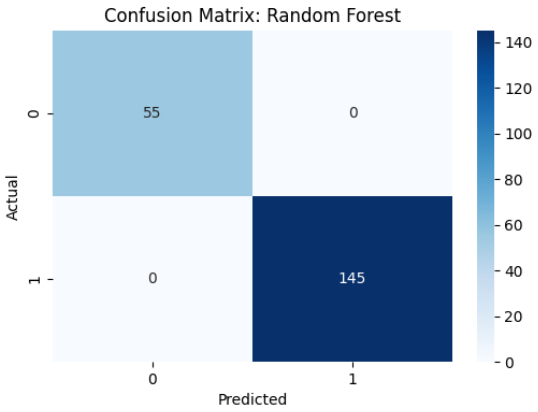

Confusion Matrix for Gaussian NB
47
8
10
135
Actual
Predicted
0
1


Confusion Matrix for K-Nearest Neighbors
51
4
5
140
Actual
Predicted
0
1


Confusion Matrix for Linear Discriminant Analysis
49
6
0
145
Actual
Predicted
0
1


Confusion Matrix for MLP Classifier
54
1
1
144
Actual
Predicted
0
1


Confusion Matrix for Support Vector Machine
0
55
0
145
Actual
Predicted
0
1


Confusion Matrix: Random Forest
55
0
0
145
Actual
Predicted
0
1

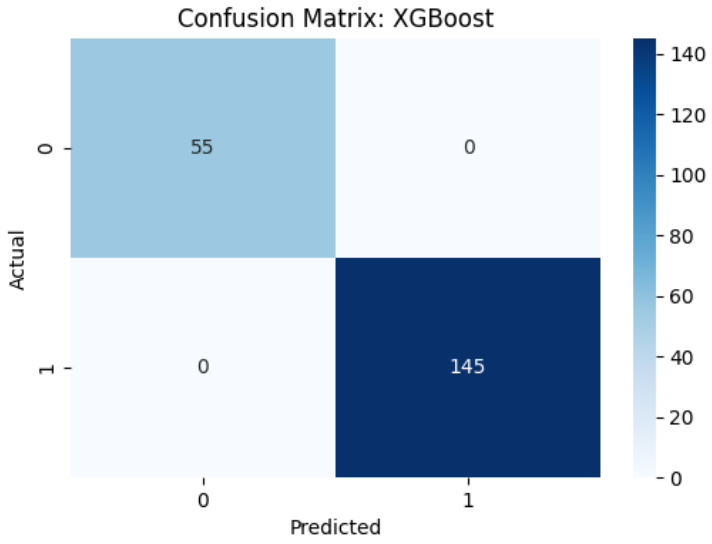


Figure 4: Confusion matrices for the evaluated classification models.

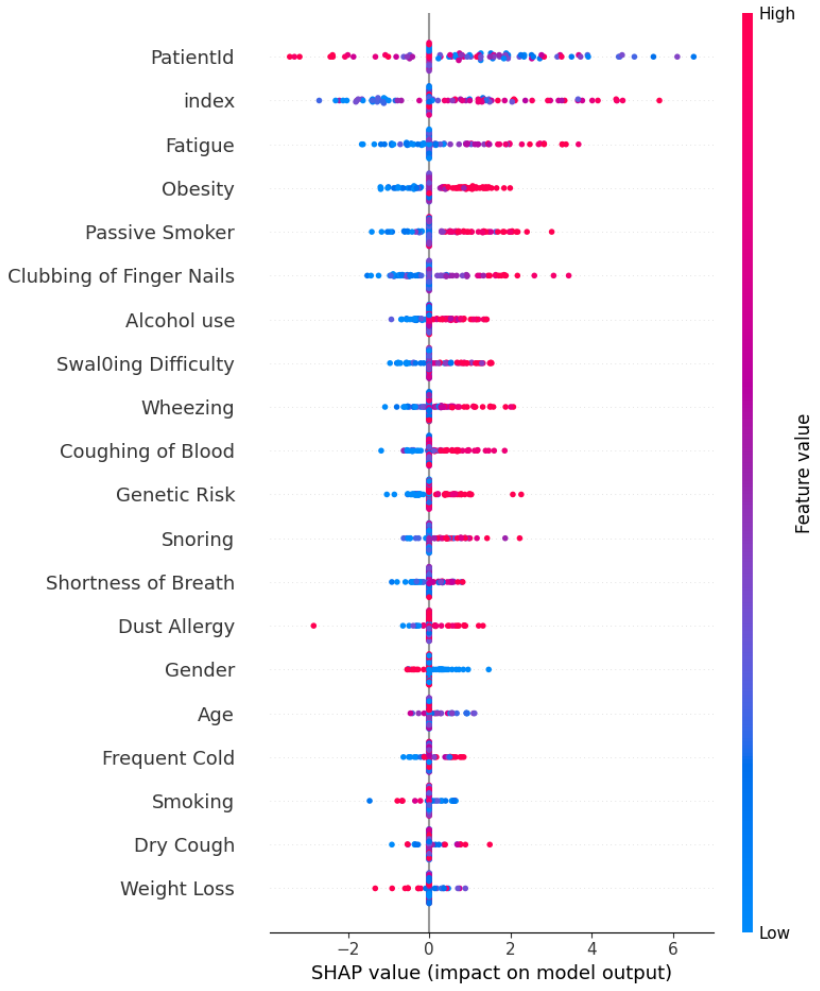


Figure 5: SHAP summary plot showing the impact of each feature on model predictions.

4.3 Feature importance analysis

To enhance model interpretability and understand the key drivers behind its predictions, SHAP (SHapley Additive Explanations) analysis was employed. The SHAP summary plot, shown in Figure 5, visualizes the impact of each feature on the model's output for a high-risk prediction. Features are ranked by their overall importance, and each point on the plot represents an individual patient, with its color indicating the feature's value (high or low) and its horizontal position showing its impact on the prediction.

The analysis reveals that Dust Allergy, Alcohol use, Fatigue, Wheezing, and Genetic Risk are the five most influential features for the model. For the most critical feature, Dust Allergy, high values (red points) are strongly associated with a positive SHAP value, significantly increasing the predicted risk of lung cancer. Conversely, low values for Alcohol use (blue points) are associated with a negative SHAP value, decreasing the expected risk. Similarly, high values for Fatigue and wheezing increase the prediction of a higher risk. This detailed, instance-level explanation provides a transparent, model-agnostic view of the decision-making process, reinforcing the clinical relevance of these factors in lung cancer risk assessment.

4.4 Computational complexity analysis

In addition to predictive accuracy, the practical utility of a clinical model depends on its computational efficiency, including training time, memory footprint, and inference speed. To evaluate this, we recorded the computational resources required for each model, as summarized in Table 3.

Table 3: Comparison of computational performance for all evaluated models.

Model	Training Time (s)	Memory Usage (MB)	Inference Time (ms/sample)
Support Vector Machine	0.3	1	0.1
K-Nearest Neighbors	0.2	0	0.4
MLP Classifier	2.3	1	0.0
Gaussian Naive Bayes	0.2	0	0.0
Linear Discriminant Analysis	0.1	0	0.0
Random Forest	0.3	0	0.1
XGBoost	0.6	5	0.2

The results indicate that traditional models, such as LDA and GNB, are the most lightweight, with minimal training times. The state-of-the-art ensemble models, XGBoost and Random Forest, required more substantial training time. Notably, while the MLP Classifier had the longest training time (2.3 seconds) due to its iterative optimization process, it achieved the fastest inference speed (0.0 milliseconds per sample). This instantaneous prediction capability makes the MLP highly suitable for real-time clinical decision support applications where rapid risk assessment is critical.

5 Discussion

Our comparative analysis provides a clear performance hierarchy for lung cancer risk classification on this dataset, as detailed in Table 4. The results unequivocally identify the tree-based ensemble methods, Random Forest and XGBoost, as the top-performing models, both achieving near-perfect F1-scores of 0.999. Their dominance is characteristic of their power in handling structured, tabular data, where they excel at modeling complex feature interactions without requiring extensive data scaling.

Despite the outstanding performance of the ensemble methods, the optimized Multilayer Perceptron classifier emerged as a highly competitive and robust alternative. With an excellent F1-score of 0.986 and perfect recall, the MLP significantly outperformed the other traditional classifiers, including Linear Discriminant Analysis, K-Nearest Neighbors, and Gaussian Naive Bayes, with statistical significance confirmed via paired t-tests ($p < 0.01$). This strong result highlights that a well-tuned neural network can closely approximate the performance of state-of-the-art classical models for this clinical task. The MLP's success was heavily reliant on a rigorous methodology, where the inclusion of dropout, L2 regularization, and hyperparameter tuning was crucial for mitigating overfitting and enhancing generalization.

The other baseline models showed varied performance. LDA also performed well with a high F1-score (0.980), while KNN was reasonable but less consistent. In stark contrast, the SVM, despite achieving perfect recall, produced an unacceptably high rate of false positives, making it unsuitable for clinical deployment.

A key contribution of this study is the enhancement of model transparency. The SHAP analysis (Figure 5) confirmed that the MLP's predictions are driven by clinically relevant features, such as Dust Allergy and Alcohol Use, providing crucial interpretability and bolstering clinical trust. This combination of high performance and explainability makes the proposed MLP-based system suitable for potential integration into EHRs as a decision support tool. This aligns with the broader trend of using deep learning for diverse health monitoring applications, ranging from sports and ergonomics risk assessment to clinical diagnostics [4].

Furthermore, while our model performed well using the available clinical variables, there is potential for future improvement through advanced feature engineering. For instance, research in cardiac care has shown that extracting nonlinear features from raw signals, such as ECGs, can significantly enhance predictive accuracy for tasks like defibrillation success [40]. A similar approach could be explored in future work by extracting more

complex, derived features from the patient data to uncover more subtle predictive patterns potentially."

Table 4: Performance comparison of the schemes drawing on evaluation metrics

Model	Precision	Recall	F1-score	F1 CI (95%)
XGBoost	0.999	0.999	0.999	[0.997–1.000]
Random Forest	0.999	0.999	0.999	[0.997–1.000]
MLP Classifier	0.973	1.0	0.986	[0.982–0.989]
Linear Discriminant Analysis	0.960	1.0	0.980	[0.975–0.984]
K-Nearest Neighbors	0.972	0.966	0.969	[0.961–0.976]
Gaussian NB	0.944	0.931	0.938	[0.927–0.948]
Support Vector Machine	0.725	1.0	0.841	[0.813–0.869]

6 Conclusion

This study conducted a rigorous comparative evaluation of an MLP for the early-stage classification of lung cancer risk. By benchmarking against a suite of models, from traditional classifiers to state-of-the-art ensembles, we demonstrated that while Random Forest and XGBoost achieved the highest performance, the optimized MLP is a powerful and highly competitive deep learning alternative, achieving an excellent F1-score of 0.986.

The novelty of this work lies in its comprehensive methodology, which combines robust preprocessing, statistical validation with confidence intervals and paired t-tests, and deep model interpretability through SHAP analysis. These steps address common limitations in prior studies and enhance the reliability and trustworthiness of the findings. The results establish the optimized MLP as a robust, transparent, and clinically relevant tool, paving the way for its use in advanced decision support systems to aid in the early detection of lung cancer.

Authors' contributions

Competing of interests

The scholars claim no competing interests.

Data availability

Data can be shared upon request.

Declarations

Not applicable

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Author statement

The manuscript has been read and approved by all the authors. The requirements for authorship, as stated earlier in this document, have been met, and each author believes that the manuscript represents honest work.

Funding

This investigation didn’t receive a specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical approval

As this study utilized a publicly available and anonymized dataset from the Kaggle repository, formal institutional review board approval was not required.

References

[1] R. Nooreldeen and H. Bach, "Current and future development in lung cancer diagnosis," *International journal of molecular sciences*, vol. 22, no. 16, p. 8661, 2021.

[2] S. Manoharan and A. Sathesh, "Early diagnosis of lung cancer with probability of malignancy calculation and automatic segmentation of lung CT scan images," *Journal of Innovative Image Processing (JIIP)*, vol. 2, no. 04, pp. 175-186, 2020.

[3] J. J. Chabon *et al.*, "Integrating genomic features for non-invasive early lung cancer detection," *Nature*, vol. 580, no. 7802, pp. 245-251, 2020.

[4] A. Aghamohammadi *et al.*, "A deep learning model for ergonomics risk assessment and sports and health monitoring in self-occluded images," *Signal, Image and Video Processing*, vol. 18, no. 2, pp. 1161-1173, 2024.

[5] M. Bhatt and P. Shende, "Advancement in machine learning: A strategic lookout from cancer identification to treatment," *Archives of Computational Methods in Engineering*, vol. 30, no. 4, pp. 2777-2792, 2023.

[6] T. Saba, "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges," *Journal of infection and public health*, vol. 13, no. 9, pp. 1274-1289, 2020.

[7] K. Pradhan and P. Chawla, "Medical Internet of things using machine learning algorithms for lung cancer detection," *Journal of Management Analytics*, vol. 7, no. 4, pp. 591-623, 2020.

[8] Z. Raeisi *et al.*, "Enhanced classification of tinnitus patients using EEG microstates and deep learning techniques," *Scientific Reports*, vol. 15, no. 1, p. 15959, 2025/05/07 2025, doi: 10.1038/s41598-025-01129-5.

[9] C. Soria, Y. Arroyo, A. M. Torres, M. Á. Redondo, C. Basar, and J. Mateo, "Method for Classifying Schizophrenia Patients Based on Machine Learning," *Journal of Clinical Medicine*, vol. 12, no. 13, p. 4375, 2023. [Online]. Available: https://www.mdpi.com/2077-0383/12/13/4375.

[10] Z. Raeisi, M. Mehrnia, R. Ahmadi Lashaki, and F. Abedi Lomer, "Enhancing schizophrenia diagnosis through deep learning: a resting-state fMRI approach," *Neural Computing and Applications*, 2025/05/22 2025, doi: 10.1007/s00521-025-11184-8.

[11] H. Tavakoli, R. Rostami, R. Shalbaf, and M. R. Nazem-Zadeh, "Diagnosis of Schizophrenia and Its Subtypes Using MRI and Machine Learning," *Brain and Behavior*, vol. 15, no. 1, p. e70219, 2025.

[12] Z. Raeisi, O. Bashiri, M. EskandariNasab, M. Arshadi, A. Golkarieh, and H. Najafzadeh, "EEG microstate biomarkers for schizophrenia: a novel approach using deep neural networks," *Cognitive Neurodynamics*, vol. 19, no. 1, p. 68, 2025/05/03 2025, doi: 10.1007/s11571-025-10251-z.

[13] C. L. Alves *et al.*, "Diagnosis of autism spectrum disorder based on functional brain networks and machine learning," *Scientific Reports*, vol. 13, no. 1, p. 8072, 2023/05/18 2023, doi: 10.1038/s41598-023-34650-6.

[14] Z. K. Khadem-Reza, R. A. Lashaki, M. A. Shahram, and H. Zare, "Automatic diagnosis of autism spectrum disorders in children through resting-state functional magnetic resonance imaging with machine vision," *Quantitative Imaging in Medicine and Surgery*, vol. 15, no. 6, pp. 4935-4946, 2025. [Online]. Available: https://qims.amegroups.org/article/view/137399.

[15] R. A. Lashaki, Z. Raeisi, A. Sodagartojgi, F. Abedi Lomer, E. Aghdaei, and H. Najafzadeh, "EEG microstate analysis in trigeminal neuralgia: identifying potential biomarkers for enhanced diagnostic accuracy," *Acta Neurologica Belgica*, 2025/05/26 2025, doi: 10.1007/s13760-025-02812-0.

[16] P. Bomatter, J. Paillard, P. Garces, J. Hipp, and D.-A. Engemann, "Machine learning of brain-specific biomarkers from EEG," *eBioMedicine*, vol. 106, 2024, doi: 10.1016/j.ebiom.2024.105259.

[17] P. C. Lee, M. W. Lin, H. C. Liao, C. Y. Lin, and P. H. Liao, "Applying machine learning to construct an association model for lung cancer and environmental hormone high-risk factors and nursing assessment reconstruction," *Journal of Nursing Scholarship*, vol. 57, no. 1, pp. 140-151, 2025.

[18] R. Ranjbarzadeh *et al.*, "Lung Infection Segmentation for COVID-19 Pneumonia Based on a Cascade Convolutional Network from CT Images," *BioMed Research International*, vol. 2021, no. 1, p. 5544742, 2021.

[19] M. K. Gould, B. Z. Huang, M. C. Tammemagi, Y. Kinar, and R. Shiff, "Machine learning for early lung cancer identification using routine clinical and laboratory data," *American journal of respiratory and critical care medicine*, vol. 204, no. 4, pp. 445-453, 2021.

[20] R. A. Lashaki, Z. Raeisi, N. Razavi, M. Goodarzi, and H. Najafzadeh, "Optimized classification of dental implants using convolutional neural networks and pre-trained models with preprocessed data," *BMC Oral Health*, vol. 25, no. 1, p. 535, 2025/04/11 2025, doi: 10.1186/s12903-025-05704-0.

[21] A. S. Moosavi, A. Mahboobi, F. Arabzadeh, N. Ramezani, H. S. Moosavi, and G. Mehrpoor, "Segmentation and classification of lungs CT-scan for detecting COVID-19 abnormalities by deep learning technique: U-Net model," (in eng), *J Family Med Prim Care*, vol. 13, no. 2, pp. 691-698, Feb 2024, doi: 10.4103/jfmpc.jfmpc_695_23.

[22] Z. Gandhi *et al.*, "Artificial intelligence and lung cancer: impact on improving patient outcomes," *Cancers*, vol. 15, no. 21, p. 5236, 2023.

[23] Q. Yuan *et al.*, "Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer," *JAMA Network Open*, vol. 4, no. 7, pp. e2114723-e2114723, 2021.

[24] P. Chaturvedi, A. Jhamb, M. Vanani, and V. Nemade, "Prediction and classification of lung cancer using machine learning techniques," in *IOP conference series: materials science and engineering*, 2021, vol. 1099, no. 1: IOP Publishing, p. 012059.

[25] V. Bulavas, V. Marcinkevičius, and J. Rumiński, "Study of Multi-Class Classification Algorithms' Performance on Highly Imbalanced Network Intrusion Datasets," *Informatica*, vol. 32, no. 3, pp. 441-475, 09/07 2021, doi: 10.15388/21-INFOR457.

[26] P. Vaitkevicius and V. Marcinkevicius, "Comparison of Classification Algorithms for Detection of Phishing Websites," *Informatica*, vol. 31, no. 1, pp. 143-160, 03/23 2020, doi: 10.15388/20-INFOR404.

[27] P. Radhika, R. A. Nair, and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms," in *2019 IEEE international conference on electrical, computer and communication technologies (ICECCT)*, 2019: IEEE, pp. 1-4.

[28] E. S. N. Joshua, M. Chakkravarthy, and D. Bhattacharyya, "An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques: A Systematic Study," *Revue d'Intelligence Artificielle*, vol. 34, no. 3, 2020.

[29] D. M. Abdullah and N. S. Ahmed, "A review of most recent lung cancer detection techniques using machine learning," *International Journal of Science and Business*, vol. 5, no. 3, pp. 159-173, 2021.

[30] Y. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Translational oncology*, vol. 14, no. 1, p. 100907, 2021.

[31] S. Huang, J. Yang, N. Shen, Q. Xu, and Q. Zhao, "Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective," in *Seminars in cancer biology*, 2023, vol. 89: Elsevier, pp. 30-37.

[32] A. Vishwakarma, A. Saini, K. Guleria, and S. Sharma, "An early prognosis of lung cancer using machine intelligence," in *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*, 2023: IEEE, pp. 1-6.

[33] V. J. Prakash and N. Karthikeyan, "Dual-layer deep ensemble techniques for classifying heart disease," *Information Technology and Control*, vol. 51, no. 1, pp. 158-179, 2022.

[34] N. Karthikeyan, "A novel attention-based cross-modal transfer learning framework for predicting cardiovascular disease," *Computers in Biology and Medicine*, vol. 170, p. 107977, 2024.

[35] A. A. V. Subramanian and J. P. Venugopal, "A deep ensemble network model for classifying and predicting breast cancer," *Computational Intelligence*, vol. 39, no. 2, pp. 258-282, 2023.

[36] T. Devastator. Lung Cancer Prediction [Online] Available: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

[37] G. A. P. Singh and P. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6863-6877, 2019.

[38] S. Huang, I. Arpaci, M. Al-Emran, S. Kılıçarslan, and M. A. Al-Sharafi, "A comparative analysis of classical machine learning and deep learning techniques for predicting lung cancer survivability," *Multimedia Tools and Applications*, vol. 82, no. 22, pp. 34183-34198, 2023.

[39] V. N. Jenipher and S. Radhika, "A study on early prediction of lung cancer using machine learning techniques," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, 2020: IEEE, pp. 911-916.

[40] H. Shamsi, A. Golkarieh, H. Nouri, and A. Dolatabadi, "Enhanced prediction of defibrillation success in out-of-hospital cardiac arrest using nonlinear ECG features and probabilistic neural network classification," *Signal, Image and Video Processing*, vol. 19, no. 8, p. 647, 2025/05/28 2025, doi: 10.1007/s11760-025-04269-3.