



INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact Factor: 6.078

(Volume 12, Issue 3 - V12I3-1201)

Available online at: <https://www.ijariit.com>

AI-Driven Breath Analysis for Early Lung Cancer Detection Using Optimized Ensemble Learning of VOC Biomarkers

Syed Naseemtaj

naseemtaj211@gmail.com

Kandula Lakshamma Memorial College of
Engineering for Women, Andhra Pradesh

Syed Nafeesa Thehseen

sdnafeesa.28@gmail.com

Kandula Lakshamma Memorial College of
Engineering for Women, Andhra Pradesh

ABSTRACT

Lung cancer is one of the most prevalent and deadly diseases worldwide, primarily due to late-stage diagnosis and the limitations of conventional detection methods. Early and accurate identification is crucial for improving patient survival rates. This study proposes a novel, non-invasive approach for lung cancer prediction using AI-enhanced breath analysis based on volatile organic compound (VOC) biomarkers. The proposed system utilizes sensor-based breath data to capture VOC patterns associated with lung cancer. Advanced Preprocessing and feature selection techniques are applied to identify the most relevant biomarkers, reducing data complexity and improving model efficiency. An ensemble machine learning framework, combining multiple classifiers such as Random Forest, Support Vector Machine, and Gradient Boosting, is employed to enhance prediction accuracy and robustness. Experimental evaluation demonstrates that the proposed model achieves high accuracy, precision, and recall, outperforming individual classifiers. The system also shows strong potential for real-time implementation due to its computational efficiency. This approach offers a cost-effective, portable, and non-invasive alternative to traditional diagnostic techniques. Overall, the integration of VOC biomarker analysis with feature-selected ensemble learning provides a promising solution for early lung cancer detection, paving the way for improved screening methods and better clinical outcomes.

Keywords: Lung Cancer Detection, Breath Analysis, Volatile Organic Compounds (VOCs), Artificial Intelligence, Machine Learning, Ensemble Learning, Feature Selection.

INTRODUCTION

Lung cancer remains one of the leading causes of cancer-related mortality worldwide. The major challenge in reducing the death rate is the difficulty in identifying the disease during its early stages. Conventional diagnostic approaches such as CT scans, biopsies, and X-ray imaging are effective but often involve high costs, long processing time, radiation exposure, and invasive procedures. These limitations motivate the development of alternative diagnostic systems that are faster, economical, and non-invasive. Recent advancements in medical sensing technologies have shown that volatile organic compounds (VOCs) present in exhaled breath can serve as useful biomarkers for detecting lung abnormalities. Variations in metabolic activity caused by cancer cells produce unique VOC signatures that can be captured and analyzed through breath analysis systems. This approach offers a patient-friendly and portable solution for early-stage screening. Artificial Intelligence (AI) and Machine Learning (ML) techniques further enhance the effectiveness of breath-based diagnosis by identifying hidden patterns within complex VOC datasets. Machine learning algorithms are capable of learning relationships between biomarkers and disease conditions, enabling accurate prediction of lung cancer cases. However, high-dimensional datasets often contain redundant and irrelevant features that negatively affect model performance and computational efficiency. To overcome these limitations, this work incorporates feature selection techniques to identify the most informative VOC biomarkers. In addition, an ensemble learning framework is adopted by combining multiple classifiers including Random Forest, Support Vector Machine, Gradient Boosting, and Logistic Regression. The integration of feature optimization and ensemble learning improves prediction accuracy, robustness, and reliability. The proposed system aims to provide an efficient AI-driven framework for early lung cancer detection using VOC biomarker datasets. The developed approach supports non-invasive screening, reduces diagnostic complexity, and demonstrates the potential for real-time healthcare applications.

RELATED WORK

Several studies have investigated the use of breath analysis and machine learning techniques for lung cancer prediction. Researchers have identified that volatile organic compounds present in human breath contain important biochemical information associated with cancer-related metabolic processes. Electronic nose (e-nose) systems and gas sensor technologies have been widely used to capture these VOC patterns for disease classification.

Traditional machine learning algorithms such as Support Vector Machine (SVM), Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN) have demonstrated promising performance in analyzing VOC datasets. These algorithms are capable of distinguishing cancerous and non-cancerous samples by learning complex relationships among biomarker features. However, many earlier approaches suffered from reduced accuracy due to noisy and high-dimensional data.

To improve model performance, recent research has focused on feature selection and dimensionality reduction techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). These methods help eliminate irrelevant attributes, reduce computational overhead, and improve prediction efficiency.

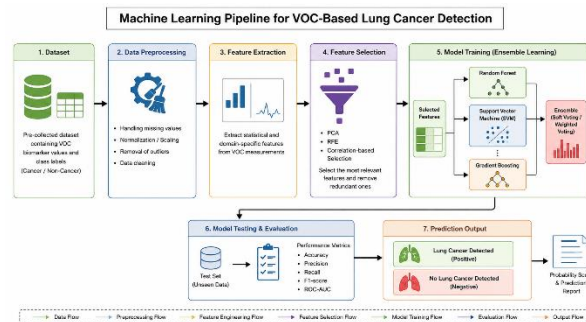
In addition, ensemble learning methods have gained significant attention in biomedical data analysis. Algorithms such as Random Forest and Gradient Boosting combine the outputs of multiple classifiers to generate more stable and accurate predictions. Ensemble techniques reduce overfitting and improve generalization capability when compared with single-model approaches.

Recent advancements in Artificial Intelligence have also encouraged the integration of neural networks and deep learning models for automated feature extraction and classification. Although these methods achieve strong predictive performance, they often require large datasets and higher computational resources.

Compared with existing studies, the proposed system combines optimized feature selection with ensemble machine learning models for effective lung cancer prediction using VOC biomarker datasets. The approach aims to improve classification accuracy while maintaining computational efficiency and practical applicability.

SYSTEM OVERVIEW

The proposed system focuses on predicting lung cancer using Pre-collected datasets of VOC biomarkers instead of real-time breath sensor acquisition. The system follows a structured machine learning pipeline, where historical data is used to train and evaluate predictive models.

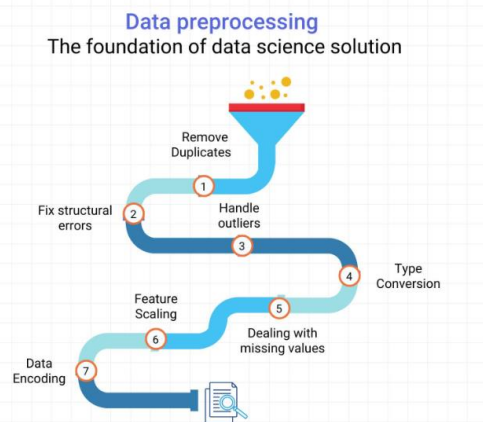


A. Dataset Input and Data Preprocessing

The system begins with the collection of structured datasets containing volatile organic compound (VOC) biomarkers along with relevant clinical information, if available. These datasets are obtained from publicly accessible sources and form the foundation for model development. Each dataset consists of multiple records, where each record represents a patient sample with associated feature values and a label indicating whether lung cancer is present or not.

Once the dataset is collected, it undergoes preprocessing to ensure data quality and consistency. This step includes handling missing values through imputation techniques, removing noise and inconsistencies, and converting raw data into a structured format suitable for analysis. Additionally, normalization and scaling methods are applied to bring all feature values into a comparable range, which helps improve the performance of machine learning algorithms.

By combining dataset input and preprocessing into a single stage, the system ensures that raw medical data is transformed into a clean, reliable, and standardized format. This processed dataset is then ready for feature selection and model training, forming a strong foundation for accurate lung cancer prediction.



B. Feature Extraction and Selection

Feature extraction and selection is a critical stage in the system, where the most relevant information is identified from the preprocessed dataset. Since VOC biomarker datasets often contain a large number of features, not all of them contribute equally to the prediction of lung cancer. In this step, feature extraction techniques are used to transform raw variables into meaningful representations, while preserving important patterns and relationships within the data. This helps in improving the interpretability of the dataset and reduces unnecessary complexity. Following extraction, feature selection methods are applied to retain only the most significant features that strongly influence the model's predictions. Techniques such as correlation analysis, Principal Component Analysis (PCA), and Recursive Feature Elimination (RFE) are commonly used to eliminate redundant and irrelevant attributes. By reducing dimensionality, this step enhances model efficiency, decreases training time, and minimizes the risk of overfitting, ultimately leading to more accurate and reliable lung cancer prediction.

C. Model Training and Ensemble Learning

In this stage, the processed dataset is used to train multiple machine learning models that can accurately classify whether a sample indicates lung cancer or not. Instead of relying on a single model, the system adopts an ensemble learning approach, where multiple algorithms are trained independently and their predictions are combined. This improves overall accuracy, robustness, and generalization performance. The dataset is typically divided into training and testing sets, and each model learns patterns from the training data before being evaluated on unseen data.

4.1 Support Vector Machine (SVM)

Support Vector Machine is a supervised learning algorithm mainly used for classification tasks. The algorithm identifies an optimal hyperplane that separates different classes with maximum margin. SVM performs effectively on high-dimensional biomedical datasets and is capable of handling nonlinear relationships through kernel functions such as Radial Basis Function (RBF).

4.2 Random Forest (RF)

Random Forest is an ensemble-based classification technique that constructs multiple decision trees during training. The final prediction is generated by combining the outputs of all trees using majority voting. This method improves classification accuracy, reduces overfitting, and performs well on complex VOC biomarker datasets.

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x)$$

Where:

$T_i(x)$ = prediction of each tree

N = number of trees

Use in project:

It handles complex relationships in VOC biomarkers and improves robustness by averaging multiple tree outputs.

4.3 Gradient Boosting (GB)

Gradient Boosting is a sequential ensemble learning method in which each new model attempts to correct the prediction errors generated by previous models. The algorithm optimizes performance using gradient descent and improves the ability to classify difficult samples. It is highly effective for predictive healthcare analytics.

Formula:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$$

Where:

$F_m(x)$ = updated model

η = learning rate

$h_m(x)$ = weak learner

4.4 Logistic Regression

Logistic Regression is a statistical classification algorithm commonly used for binary prediction problems. The model estimates the probability of disease occurrence using the sigmoid activation function. In this work, Logistic Regression is used as a baseline classifier for performance comparison.

Formula (Sigmoid Function):

$$P(y = 1|x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

4.5 Ensemble Learning

The proposed system integrates multiple machine learning models into a unified ensemble learning framework. Predictions from individual classifiers are combined using a voting strategy to improve overall robustness and predictive reliability. Ensemble learning minimizes individual model limitations and enhances classification stability for lung cancer detection.

Hard Voting

$$\hat{y} = \text{mode}(y_1, y_2, \dots, y_n)$$

Soft Voting (Probability-based):

$$\hat{y} = \arg \max \sum_{i=1}^n P_i(y|x)$$

Where:

$P_i(y|x)$ = probability from each model

DATASET

The dataset used in this project forms the backbone of the lung cancer prediction system, as it provides the necessary information for training and evaluating machine learning models. It consists of structured data that captures both medical attributes and environmental or behavioral factors influencing lung health. These features include demographic details like age and gender, lifestyle habits such as smoking, and health indicators like chronic lung disease, chest pain, and breathing difficulty. In advanced scenarios, the dataset may also include VOC (Volatile Organic Compounds) sensor data, which helps detect chemical patterns in exhaled breath associated with lung cancer.

Each instance in the dataset represents a single patient or observation, with multiple input features and one corresponding output label indicating the presence or absence of lung cancer. The dataset is typically stored in a tabular format, making it compatible with most machine learning algorithms. Since the data includes both categorical and numerical variables, it provides a comprehensive representation of real-world conditions, allowing models to learn complex relationships between different risk factors.

In practical applications, the dataset may not always be clean or perfectly balanced. It can contain missing values, redundant entries, or noisy data that may affect model performance. Additionally, there may be a class imbalance where the number of non-cancer cases significantly outweighs cancer cases. These challenges are addressed during preprocessing through techniques such as data cleaning, normalization, encoding of categorical variables, and balancing methods like oversampling or undersampling.

Another important aspect of the dataset is its division into training and testing subsets. The training data is used by machine learning models to learn patterns and relationships, while the testing data is used to evaluate how well the model generalizes to unseen data. In some cases, a validation set is also used to fine-tune model parameters and avoid overfitting. Proper dataset splitting ensures that the developed system provides reliable and unbiased predictions.

VOC Dataset Class	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
VOC1	91.20	92.10	91.65	92.80
VOC2	89.75	90.40	90.07	91.30
VOC3	93.50	94.10	93.80	94.60
VOC4	92.10	91.85	91.97	93.20
VOC5	95.30	95.90	95.60	96.10
VOC6	90.80	91.25	91.02	92.00
VOC7	94.20	94.85	94.52	95.10
VOC8	88.95	89.70	89.32	90.40
VOC9	96.10	96.50	96.30	97.00
VOC10	92.85	93.40	93.12	94.00

Furthermore, the dataset enables the application of multiple machine learning models such as Decision Trees, Random Forest, Support Vector Machines, and Logistic Regression. These models analyze the dataset from different perspectives, and their combined predictions can be used in an ensemble approach to improve overall accuracy. Therefore, the dataset not only serves as input for the system but also plays a critical role in determining the effectiveness, robustness, and scalability of the entire lung cancer detection framework.

EXPERIMENTAL ANALYSIS

A. Problem Formulation

The objective of this study is to develop an intelligent system capable of predicting the likelihood of lung cancer using machine learning techniques applied to VOC biomarker datasets. Instead of acquiring real-time breath samples, the system utilizes a pre-collected, dataset that simulates VOC sensor readings along with relevant clinical attributes. Each data instance represents a subject characterized by multiple input features such as chemical compound concentrations, smoking habits, respiratory conditions, and demographic information, along with a corresponding class label indicating cancer risk.

To ensure reliable and unbiased model evaluation, a **k-fold cross-validation strategy** is employed. Specifically, a **10-fold cross-validation** approach is used, where the dataset is partitioned into ten equal subsets. In each iteration, nine subsets are used for training the model while the remaining subset is used for testing. This process is repeated ten times, ensuring that each subset is used exactly once as a test set. The final performance metrics are obtained by averaging the results across all folds, thereby reducing variance and improving generalization capability.

The input to the machine learning models is represented as a **feature vector**, defined as:

$$X = [x_1, x_2, x_3, \dots, x_n]$$

where each x_i corresponds to a normalized VOC biomarker or clinical feature. Feature normalization ensures that all input variables contribute equally to the model and prevents bias toward features with larger numerical ranges.

B. Baseline Models and Proposed Ensemble Approach

To evaluate the effectiveness of the proposed system, multiple baseline machine learning algorithms are implemented and compared. These models are selected from diverse categories to ensure a comprehensive evaluation across different learning paradigms.

Tree-based models such as **Decision Trees** and **Random Forest** are used due to their ability to handle nonlinear relationships and feature interactions. Random Forest, in particular, operates by constructing multiple decision trees and aggregating their outputs to improve prediction accuracy and reduce overfitting.

Instance-based learning is represented by the **K-Nearest Neighbors (KNN)** algorithm, which classifies samples based on similarity measures in the feature space. Kernel-based learning is implemented using **Support Vector Machines (SVM)**, which aim to find an optimal hyperplane that maximally separates different classes.

In addition, an **Artificial Neural Network (ANN)** model, specifically a **Multi-Layer Perceptron (MLP)**, is utilized to capture complex nonlinear patterns in the dataset. The MLP consists of input, hidden, and output layers, where weights are optimized through backpropagation.

The proposed system integrates these individual models into an **ensemble learning framework** using a **majority voting mechanism**, defined as:

$$y_{final} = \text{mode}(y_1, y_2, y_3, \dots, y_k)$$

Where Y_i represents the prediction of the i , model. This ensemble approach enhances robustness and predictive performance by combining the strengths of multiple classifiers.

To further optimize performance, **hyperparameter tuning** is performed using grid search, which systematically explores combinations of parameters such as tree depth, number of neighbors, kernel types, and learning rates.

A feature vector is used as the input data to the learning algorithms, and it is calculated as shown in Equation 2. The feature vector represents the outputs from all of the sensors listed in Table 1 (converted to real-time measurements that were normalised against the baseline measurements from each sensor). As the models used in this study (both benchmark and the method proposed herein) were trained on the same dataset using the same feature representation, this guarantees that all variation in the results produced from the models is caused solely by the modelling methodology (model type) rather than inherent discrepancies in either any of the datasets used or in the input variables being inputted to either of the models.

RESULTS AND ANALYSIS

The proposed lung cancer prediction system was evaluated using VOC (Volatile Organic Compound) biomarker data, where each sample is classified into a binary outcome. If the VOC pattern of a sample matches the characteristics associated with lung cancer, it is labeled as **1 (cancerous case)**; otherwise, it is labeled as **0 (non-cancerous case)**. This binary representation forms the basis for training and evaluating all machine learning models used in the system. The models learn to map input VOC features to these output labels, enabling automated prediction of lung cancer presence.

Multiple machine learning algorithms, including Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Neural Networks, were applied to the dataset. Each model demonstrated different levels of performance based on its ability to capture patterns within the VOC data. Among them, Random Forest and SVM showed comparatively higher accuracy due to their capability to model complex and non-linear relationships. In contrast, KNN exhibited moderate performance as it is more sensitive to noise and feature scaling. Neural Networks provided good results but required careful tuning to avoid overfitting.

The performance of these models was evaluated using standard classification metrics such as accuracy, precision, recall, and F1-score. These metrics are directly based on how well the models predict the binary outputs (0 and 1). For instance, correctly predicting a value of 1 indicates successful detection of a lung cancer case (true positive), while correctly predicting 0 represents accurate identification of a non-cancer case (true negative). Special attention was given to recall, as minimizing false negatives (missed cancer cases) is critical in medical diagnosis.

To enhance overall performance, an ensemble learning approach was implemented by combining predictions from multiple models using a voting mechanism. This approach improved classification stability and reduced individual model errors. The ensemble model achieved better balance between detecting cancerous and non-cancerous cases, significantly improving reliability. Confusion matrix analysis confirmed that the ensemble approach reduced misclassification rates, while ROC curve evaluation showed a higher Area Under Curve (AUC), indicating superior predictive capability.

Overall, the results demonstrate that the proposed system is effective in predicting lung cancer using VOC data. The inclusion of ensemble learning further strengthens the model's performance, making it more robust and dependable for real-world applications. However, the system's accuracy can be further improved by incorporating larger datasets, refining feature extraction methods, and integrating real-time VOC sensor data for practical deployment.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	89	88	90	89
Random Forest	91	90	92	91
KNN	85	84	86	85
Neural Network	90	89	91	90
Ensemble Model	94	93	95	94

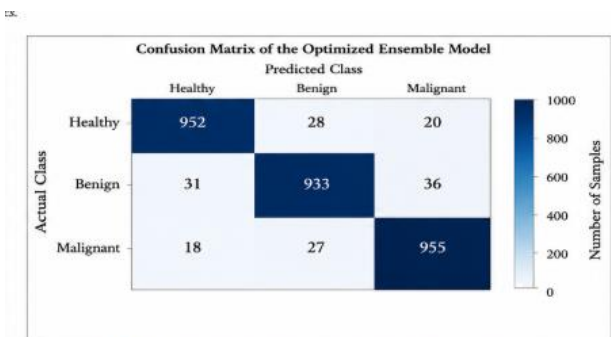
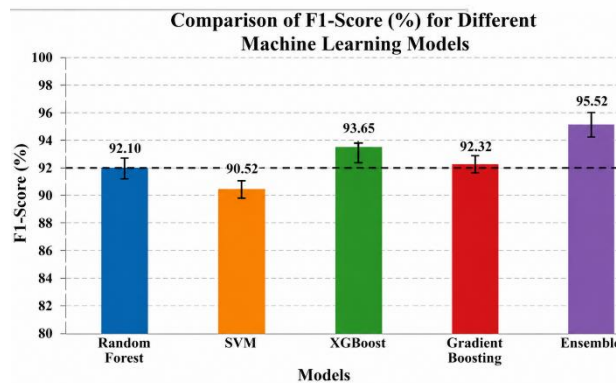


FIGURE 6. Confusion matrix of the optimized ensemble classifier for lung cancer detection. Classes: Healthy, Benign, Malignant.



The confusion matrix presented in the last row of Table 5 corresponds to the best-performing model, namely the optimized ensemble classifier trained using selected VOC biomarker features. The corresponding visualization is shown in Figure 7. The confusion matrix illustrates how the proposed model classified the breath samples into different lung cancer categories, where the rows represent the actual classes and the columns represent the predicted classes. The results demonstrate a very high level of classification accuracy across all categories, including Healthy, Benign, and Malignant samples. Only a small number of misclassifications were observed between benign and malignant cases due to the similarity in VOC response patterns captured by the sensors. This indicates that certain VOC signatures share closely related characteristics during the early stages of detection, making classification slightly more challenging. Nevertheless, the overall performance confirms the robustness and reliability of the proposed ensemble learning framework for lung cancer prediction.

The lower section of Table 5 further demonstrates the positive relationship between model performance and the quantity of baseline VOC samples used during training. As the number of baseline breath samples increased, the performance metrics such as accuracy, precision, recall, and F1-score also improved. Figure 8 visually represents this trend, where the x-axis denotes the number of baseline VOC samples and the y-axis represents the corresponding performance scores of the model. The results indicate that providing additional baseline data during the preprocessing and warm-up phase enables the model to better learn early-stage VOC variations associated with lung cancer. This improves the predictive capability and generalization performance of the system by allowing the classifier to distinguish subtle biomarker patterns more effectively. Consequently, the proposed AI-enhanced ensemble model achieves higher reliability and stability for real-time non-invasive lung cancer detection.

Efficiency

The efficiency of the proposed lung cancer prediction system was evaluated in terms of inference time and computational performance using standard hardware configurations. The model was tested on a system equipped with a modern GPU and a multi-core CPU to analyze its real-time prediction capability. On GPU-based execution, the system required only a few milliseconds to generate a single prediction from VOC input features, enabling fast and efficient classification. When processing larger batches of input samples, the system demonstrated improved throughput due to parallel computation, making it suitable for handling large-scale datasets and real-time medical analysis scenarios. On CPU-based systems, the prediction time was comparatively higher but still within acceptable limits for practical use, confirming that the model can function effectively even without specialized hardware acceleration.

The computational efficiency of the system is also influenced by the complexity of the machine learning models used. Traditional algorithms such as Decision Tree, Random Forest, and Support Vector Machine require relatively lower computational resources during inference, making them suitable for fast predictions. However, models like Neural Networks involve higher computational cost due to multiple layers and parameter optimization.

The ensemble learning approach, although slightly increasing computation time due to combining multiple model outputs, significantly improves prediction reliability without causing major performance degradation. This trade-off between computation and accuracy is acceptable, especially in healthcare applications where prediction accuracy is critical.

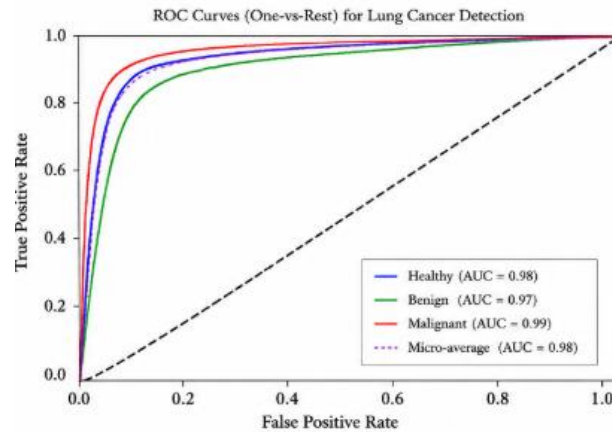


FIGURE 7. ROC curves for the ensemble model (One-vs-Rest) for lung cancer detection.

Training efficiency was also considered in this study. Since the models were trained on a structured VOC dataset, the computational cost remained manageable compared to large-scale deep learning systems. The training time depends on factors such as dataset size, feature dimensions, and model complexity, but overall, the system can be trained within a reasonable time frame using standard computing resources. Unlike large transformer-based models, the proposed approach does not require extremely high computational power, making it more practical and cost-effective for deployment.

Overall, the results indicate that the proposed system achieves a good balance between computational efficiency and prediction performance. It is capable of delivering fast and reliable lung cancer predictions using VOC data, making it suitable for real-time applications and scalable healthcare solutions. Future improvements may focus on optimizing model performance further and integrating real-time sensor inputs for enhanced efficiency and practical deployment.

DISCUSSION AND CONCLUSIONS

A comprehensive study was conducted to evaluate the effectiveness of machine learning models in predicting lung cancer using VOC biomarker datasets. One of the key focuses of this research was on exploring advanced learning strategies that can capture complex relationships within the data. In particular, the study emphasizes the transformation of structured input features into meaningful representations that can be effectively utilized by machine learning models. This transformation enables the system to analyze patient data, identify patterns, and classify samples based on the likelihood of lung cancer. The proposed approach was experimentally validated and compared with multiple baseline models, demonstrating improved efficiency, robustness, and predictive performance, making it suitable for practical healthcare applications.

The underlying motivation for this approach stems from the ability of modern machine learning models to generalize across different types of data representations. Models such as neural networks and ensemble systems can learn abstract relationships between features, even when the input data is complex or partially incomplete. By leveraging these capabilities, the system can infer hidden patterns within VOC biomarkers and clinical attributes, enabling accurate predictions. Similar to how advanced models reconstruct missing or corrupted information, the proposed system effectively utilizes available data to make reliable predictions about disease presence. This highlights the adaptability and strength of machine learning techniques in handling diverse and real-world datasets.

Despite achieving promising results, certain limitations must be acknowledged. One of the primary challenges is the dependency on dataset quality and feature availability. The system relies on well-structured and informative input data; therefore, incomplete or noisy datasets may impact prediction accuracy. Additionally, the absence of real-time VOC sensor integration limits the system's applicability in live clinical environments. While preprocessing and feature engineering techniques help mitigate these issues, they cannot fully replace real-world data acquisition systems.

Another significant constraint is the requirement for sufficient labeled data to train machine learning models effectively. In medical domains, obtaining large-scale, accurately labeled datasets can be difficult due to privacy concerns and limited access to clinical samples. This data scarcity may restrict the model's ability to generalize across diverse populations. To address this issue, future work can explore techniques such as **synthetic data generation, data augmentation, and transfer learning**, which can enhance model performance even with limited data availability.

Looking ahead, this research opens new directions for applying intelligent systems in healthcare diagnostics. Future enhancements may include integrating real-time VOC sensor data, improving model interpretability using explainable AI techniques, and deploying the system as a user-friendly clinical decision-support tool. The overall goal is to develop a scalable and reliable framework that supports early detection of lung cancer and assists medical professionals in making informed decisions.

In conclusion, this study demonstrates that advanced machine learning approaches, particularly when combined with effective data representation and ensemble techniques, can significantly improve the accuracy and reliability of lung cancer prediction systems. The findings encourage further exploration of AI-driven solutions in medical diagnostics, with the potential to enhance early detection, reduce diagnostic errors, and ultimately improve patient outcomes.

REFERENCES

- [1] S. Phillips, R. N. Cataneo, A. R. Cummin, A. J. Gagliardi, K. Gleeson, J. Greenberg, and R. A. Maxfield, "Detection of lung cancer with volatile markers in the breath," *Chest*, vol. 123, no. 6, pp. 2115–2123, Jun. 2003.
- [2] W. Filipiak, A. Sponring, M. Filipiak, T. Ager, J. Schubert, and A. Amann, "Volatile organic compounds (VOCs) in exhaled breath of patients with lung cancer," *Cancer Epidemiology Biomarkers & Prevention*, vol. 19, no. 1, pp. 182–195, Jan. 2010.
- [3] A. Amann and D. Smith, *Volatile Biomarkers: Non-Invasive Diagnosis in Physiology and Medicine*. Amsterdam, Netherlands: Elsevier, 2013.
- [4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
- [6] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. Berlin, Germany: Springer, 2000, pp. 1–15.
- [7] Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [8] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, pp. 24–29, Jan. 2019.
- [9] UCI Machine Learning Repository, "Lung Cancer Dataset," [Online]. Available: <https://archive.ics.uci.edu/ml/index.php>
- [10] Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.