RESEARCH ARTICLE

AI-Driven Predictive Modeling for Lung Cancer Detection and Management Using Synthetic Data Augmentation and Random Forest Classifier

Nisreen Innab[1] · Asma Aldrees[2] · Dina Abdulaziz AlHammadi[3] · Abeer Hakeem[4] · Muhammad Umer[5] · Shtwai Alsubai[6] · Silvia Trelova[7] · Imran Ashraf[8]

Received: 3 January 2025 / Revised: 31 March 2025 / Accepted: 18 May 2025

Abstract
Artificial intelligence (AI) transforms multiple businesses, including medical research, where AI-driven developments bring significant advantages. The application of machine learning algorithms enables medical researchers to examine large amounts of data accurately, which leads to the development of precise and effective treatment approaches. Lung cancer leads the list of critical healthcare issues because it remains the world's most lethal form of cancer thus demanding innovative diagnostic tools for faster and accurate identification. The proposed study introduces an innovative method called CTGAN-RF which uses conditional tabular generative adversarial networks (CTGAN) and random forest (RF) classifier to detect lung cancer through synthetic data generation. The proposed model demonstrated superior performance by achieving a 0.9893 score of accuracy and 0.99 value for precision, F1 score, and recall. Extensive experimental evaluation for this method included testing nine classification algorithms. The implementation of different classifiers employed data balancing methods including SMOTE and borderline-SMOTE along with SMOTE ENN and unbalanced data configurations. Comparative analysis showed that CTGAN-RF consistently performed significantly better than traditional classifiers in dealing with class imbalance and improving prediction accuracy. After testing with fivefold cross-validation, the reliability of the model was further validated. In comparison to cutting-edge approaches for lung cancer diagnosis, the proposed methodology outperformed in terms of classification metrics. This in-depth evaluation of synthetic data augmentation with machine learning in lung cancer detection has helped in the development of personalized treatment strategies in the fight against such a life-threatening disease.

Keywords Lung cancer detection · Data augmentation · Machine learning in healthcare · Healthcare · CTGAN-RF model · Class imbalance handling

1 Introduction

The advent of artificial intelligence (AI) has revolutionized several sectors, such as engineering, manufacturing, and healthcare. Specifically, the integration of machine learning (ML) techniques into medical science is significantly advancing the field through the development of automatic and treatment-based solutions. A major role of these models is to help healthcare professionals create precise, reliable, and personal treatment strategies based on processing medical data with great efficiency. Lung cancer is one of the most critical types of cancer and causes a high rate of mortality throughout the world. Early detection and intervention are important because they can result in great improvement in patient survival rates and reduce mortality. Over 40.9 million people were living with

Published online: 10 June 2025

lung cancer in 2017 (causing a subsequent DALY rate of 37.6 DALYs per 100,000 population). It was estimated to cause around 1.8 million deaths worldwide in 2020 [1, 2]; and in the United States, there were projected to be 135,720 deaths. One of the major problems in lung cancer fighting is that this disease is non-symptomatic in the early stages and is quite difficult to detect over time; consequently, diagnosis is hard and complex. It is based on the ability of approaches for analyzing and predicting the progression and development of lung cancer [3]. The body's most aggressive type of cancer, accounting for 10 to 16 percent of all lung cancer, is small cell lung cancer (SCLC). SCLC is highly aggressive, with tumors that rapidly develop and metastasize, accelerating the disease's progression. Smoking is a strong causal factor for this type of cancer, reinforcing the need for better predictive models to treat its growth and spread. This study uses AI-driven methodologies to help with earlier detection and better management of lung cancer which will improve patient outcomes.

Recent advancements in deep learning have led to the development of sophisticated models for medical image analysis [4, 5] and disease classification [6, 7]. For instance, Ahamed et al. [8] introduced a MultiResUNet framework with attention-guided segmentation for the automated detection of colorectal polyps from endoscopic images. In another study, a collaborative federated learning framework was proposed for lung and colon cancer classifications [9], emphasizing data privacy while maintaining high performance. Furthermore, a Multi-Scale CNN integrated with Explainable AI [10] techniques was developed to enhance the classification of lung-affected diseases, providing interpretable insights into model decisions. Additionally, the IRv2-Net model [11] combined InceptionResNetV2 and UNet architectures with test-time augmentation techniques to improve polyp segmentation performance.

ML can predict the likelihood that a patient will develop lung cancer shortly by identifying patterns and risk factors associated with the disease through the analysis of extensive patient records. Advances in bioinformatics, particularly in image and factor-based analysis, have enabled the early detection of lung cancer. The core idea of this approach lies in determining the patient's risk by benefiting from the knowledge of those key features in those datasets and developing a reliable ML model. However, such algorithms vastly improve malignancy prediction accuracy. To assess the performance of the various metrics used: confusion matrix, precision, F1 score, recall, and accuracy. Several ML methods are used to create gene expression patterns from samples and build predictive models for cancer detection [12].

The authors developed and assessed an ADB-based prognostic model to assist healthcare providers in making ECOG PS-informed decisions for lung cancer patients. Numerous studies have explored the application of ML in this specific area of medicine. This study introduces a framework for creating efficient ML classification models to predict lung cancer, using everyday routines and clinical symptoms/signs as input features. These days, AI and ML support the healthcare sector. ML models' effectiveness is often related to models' ability to classify true or false positives and negatives in various datasets [13]. The findings of this study demonstrate notable advancements compared to prior research [14, 15], particularly in evaluating the performance of various classifiers such as XGBoost, Extra Trees Classifier (ETC), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Stochastic Gradient Descent Classifier (SGDC), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boosting Classifier (GBC). These classifiers were analyzed for their ability to detect lung cancer at an early stage by examining relevant medical data [16].

This research examines different methods that scientists have applied when predicting lung cancer based on clinical risk factor measurements. This work introduces different approaches for survival prediction along with strategies to select features and extract them and it demonstrates how ML and DL techniques operate through multiple validation evaluation metrics. Several researchers developed a manually created electronic nose (e-nose) device as a detection system for lung cancer [17, 18]. Significant advances have been designed but future developments must focus on building better classification methods in the field. A wide range of researchers have adopted feedforward and backpropagation approaches in neural networks for better classification accuracy [19, 20]. Research teams have concentrated on AI-driven detection methods for diseases alongside medical decision-making for an extended timeframe. The main study findings presented the following points:

- Introduced an innovative framework for classifying lung cancer by combining CTGAN-generated synthetic features with a Random Forest classifier, setting a new benchmark in disease classification.
- Attained remarkable performance metrics, achieving a 0.9893 score of accuracy, along with precision, F1 score, and recall rates of 0.99, showcasing substantial advancements over conventional approaches.
- Performed an in-depth comparative analysis against nine established machine learning classifiers, including XGBoost, Extra Trees Classifier (ETC), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Stochastic Gradient Descent Classifier (SGDC), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Gradient Boosting Classifier (GBC). The evaluation incorporated SMOTE, Borderline-SMOTE, SMOTE ENN, and original datasets to ensure a thorough and balanced assessment.
- Validated the proposed model's reliability and robustness through 5-fold cross-validation, outperforming earlier studies and establishing its credibility.
- Demonstrated the potential of combining advanced synthetic data generation methods with machine learning models to enhance early lung cancer detection and prediction, paving the way for next-generation diagnostic tools in healthcare.

The paper is organized as follows: Section 2 provides a comprehensive review of existing literature in the field of lung cancer prediction, highlighting key findings from prior studies. Section 3 outlines the methodology employed in this research, including details about the dataset, the proposed framework, descriptions of the machine learning models used, and the evaluation metrics applied. Section 4 presents the experimental setup, results, and a detailed analysis of each learning model under various scenarios explored in this study. Finally, Section 5 concludes the paper by summarizing the findings and their implications.

2 Literature Review

Early detection and accurate prediction of lung cancer have become an important research area since high rates of mortality accompany such cancer and demand timely diagnosis. During the past few years, enormous improvements have been achieved in ML and DL techniques, which have opened doors to novel methods to solve these problems. Clinical data, easily accessible imaging data as well as demographic information are being explored by researchers worldwide to improve predictive models. This section briefly reviews the recent studies benefitting the field with the application of ML algorithms for extracting features and generating synthetic data to improve lung cancer risk prediction. Based on these works, we position the current state of research and delineate gaps that we are looking to address with our proposed framework.

The researchers [21] investigated the use of ML and predictive modeling for enhanced public health system detection of lung cancer in early stages through their research. The research method aimed to decrease medical expenses through precise forecasting that enabled personalized care plans leading to reduced unnecessary medical treatment costs. The authors examined five machine learning applications namely LR, XGBoost, SVM, LightGBM, and AdaBoost. The XGBoost algorithm demonstrated the best results with 96.92% accuracy when used as the model.

Similarly in another work, the authors [22] studied the use of ML models to predict lung cancer risk. The researchers developed models to detect high-risk patients while initiating prompt treatments, which would help minimize the long-term effects of illnesses. The research team built a Rotation Forest model which they evaluated through important performance metrics that included precision, accuracy, recall, F1 score, and AUC. Their results showed remarkable performance, with an AUC of 99.3% and an F score, recall, precision, and accuracy of 97.1%.

The researchers [23] investigated textual data and ML algorithms for lung cancer prediction. They used the University of California cancer dataset to develop an SVM model that improved diagnosis accuracy. They compared their unique methodology with classic SVM and SMOTE algorithms, and it achieved an impressive 98.8% accuracy rate. These studies illustrate ML's potential for improving early lung cancer identification and patient outcomes.

The researchers [24] used ML techniques to detect lung cancer risks through analysis of textual patient data. The authors stressed the necessity of retrieving medical and demographic information from medical records followed by dataset cleaning operations before implementing training for ML models. The researchers worked to build a dependable practical ML system that detects lung cancer at its early stages through the analysis of demographic and clinical information. A predictive model with maximum accuracy was recommended by the team through the implementation of LR together with DT and RF, SVM, KNN, and NB.

The researchers [25] investigated ML techniques applied for the prediction of lung cancer. The researchers examined kNN, LR, SVM, and GNB as different ML approaches. The research dataset came from Kaggle for their investigation. Their experimental models demonstrated that their designed approach achieved better outcomes compared to other evaluation parameters which established it as the leading method for lung cancer prediction. The combination of research results demonstrates the ability of ML to improve both the early detection and risk prediction of lung cancer.

The authors of [26] used XGBoost and KNN to perform lung cancer predictions. The research utilized XGBoost and KNN as ML algorithms to forecast cancer risks. The accuracy rates alongside precision and recall values reached equal levels for both testing models. The complex nature of the dataset led to XGBoost producing better results than KNN when identifying advanced data patterns. The researchers [27] introduced an optimization technique for lung cancer classification based on hyperparameter tuning. Their research evaluated and compared four strategies against existing methods using datasets sourced from Kaggle. The proposed method worked to fine-tune both Gamma and C parameters because Gamma establishes kernel width and C regulates the amount of regularization used in the algorithm. Their method proved to be effective based on the results achieved, which measured 99.16% accuracy 98% precision, and 100% sensitivity. Rani et al. [28] developed an ML-based risk prediction model to identify disease trends and to improve early detection of lung cancer using disease-related data. The authors compared the performance of NB and SVM algorithms for lung cancer classification. Together, these studies have shown the efficacy of ML in enhancing lung cancer prediction and risk assessment.

The researchers [29] proposed an ML method to improve the classification accuracy of lung cancer using image-sized numerical data. In this research work, extensive preprocessing improved the classification results of the proposed model to a notable extent. The KNN algorithm achieved an outstanding accuracy score of 87% by utilizing PCA features. In another research work [30], authors make use of six machine learning models for diagnosing lung cancer with numerical and categorical data. Out of six ML models, SVM performed well with 91% accuracy, but still lower than other research works that make use of data augmentation techniques to resolve the class-imbalance problem.

Early detection and accurate prediction of lung cancer have become important areas of research because this type of cancer has a high death rate and needs to be diagnosed quickly. During the past few years, enormous improvements have been achieved in ML techniques which have opened doors to novel methods to solve these problems. Clinical data, easily accessible imaging data as well as demographic information are being explored by researchers worldwide to improve predictive models. This section briefly reviews the recent studies benefitting the field with the application of ML algorithms for extracting features and generating synthetic data to improve lung cancer risk prediction. This work seeks to contribute to the ongoing efforts in leveraging AI-driven solutions to transform lung cancer diagnosis and patient outcomes (Table 1).

3 Materials and Methods

The proposed methodology, along with its step-by-step workflow towards early identification of lung cancer is explained in Fig. 1. This process includes getting the dataset, doing the preprocessing, splitting the data into training and testing sets, applying the oversampling techniques, and training the models. Finally, model predictive accuracy and robustness in predicting outcomes are evaluated.

Table 1 Related work summary

References	Classifiers	Dataset	Performance
[21]	XGBoost, AdaBoost, LightGBM, Logistic Regression, and Support Vector Machine	Dataset from hospitals situated in Dhaka, Bangladesh 5000 instances	XGBoost 96.92%
[22]	NB, BayesNet LMT, RT, LR, RF, ANN, SGD, SVM, DT(RepTree), 3NN, J48, RotF, AdaBoostM1	Kaggle 309 instances	RF 97.1%
[23]	KNN, Naive Bayes, SVM, J48	University of California Lung Cancer Dataset: 32 instances	SVM with SMOTE 98.8%
[24]	LR, RF, DT, KNN, SVM, NB		RF 90.32%
[25]	KNN, SVM, LR, Naive Bayes	Kaggle 53,428 instances	SVM 98%
[26]	KNN and XGBoost	Kaggle 1000 instances	XGBoost 100%
[27]	SVM, DT, XGB, and logistic regression (LR)	Kaggle 309 instances	SVM 99.16%
[28]	Naive Bayes and SVM	Dataset from Data World Repository: 10,531 instances	SVM 96%
[29]	RF, KNN, NB, LR, DT, SVM	UCI	KNN 87% with PCA features
[30]	NB, LR, DT, RF, GB, and SVM	Kaggle 309 instances	91% SVM

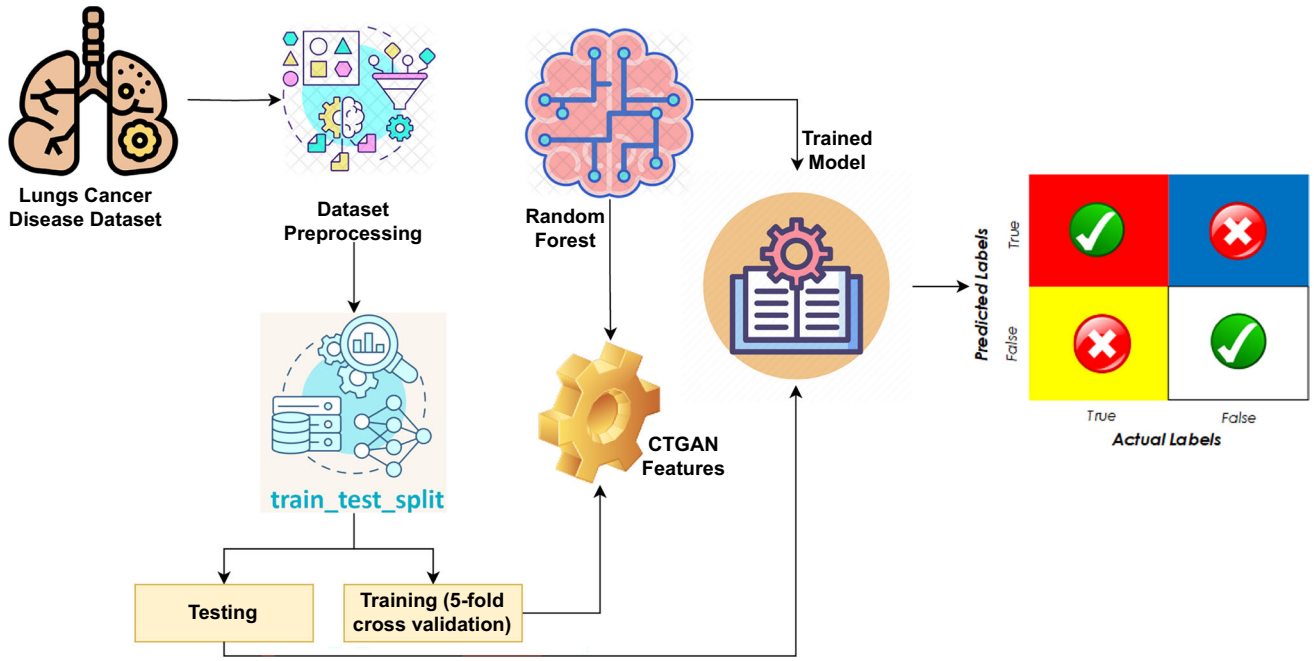

Fig. 1 Workflow of the proposed framework

3.1 Dataset

In this study, the dataset was obtained from the public repository Kaggle [31]. It has 16 features (15 of which are predictive attributes and 1 is the class attribute) of which their entries total 309. The presence or absence of Lung Cancer is given in the class attribute and factors like Gender, Age, Lung Cancer, Anxiety, Smoking, Peer Pressure, Yellow Fingers, Chronic Disease, Allergy, Fatigue, Alcohol Consumption, Wheezing, Coughing, Swallowing Difficulty, Shortness of Breath, Chest Pain are taken into account as predictive attributes. It is found that the class imbalance in the dataset is quite large, with only 39 occurrences being labeled "normal," compared to 270 "cancer."

 https://doi.org/10.1007/s44196-025-00879-4

Table 2 Dataset statistics

Attribute	Description
Gender	Specifies individual gender(M or F)
Age	Records individual's age
Anxiety	Indicates if the individual experiences anxiousness
Smoking	Specifies if the individual is a smoker
Peer pressure	Indicates the individual's sensitivity to peer pressure
Yellow fingers	Specifies individual yellow fingers (yes or no)
Chronic disease	Specifies individual chronic disease presence (yes or no)
Allergy	Shows whether the individual has allergies
Fatigue	Indicates the level of weariness in the individual
Alcohol	Specifies if the individual consumes alcohol
Wheezing	Indicates if the individual experiences wheezing
Coughing	Shows if the individual has a cough
Swallowing difficulty	Indicates if the individual has trouble swallowing
Lung cancer	Specifies individual has lung cancer (yes or no)
Shortness of breath	Indicates if the individual experiences shortness of breath
Chest pain	Shows if the individual experiences chest pain

Table 2 provides a detailed breakdown of the characteristics of the dataset.

3.2 Pre-processing

Preprocessing data is essential in improving the performance of the ML algorithms. In this stage, I converted categorical attributes into numerical values (0 and 1) for the training of the model. Problems such as noise, missing values, and high imbalance in data lead to low prediction accuracy. Although there was no missing value in the dataset, it was extremely imbalanced, with 270 instances placed in the cancer class and 39 placed in the no cancer class. Since this imbalance was present, we mitigate it by applying techniques such as SMOTE, Borderline SMOTE, SMOTE-ENN, and CTGAN. While the dataset utilized in this study had no missing values, it was extremely skewed, with 270 cases categorized as "cancer" and just 39 as "no cancer." To address this imbalance issue approaches such as SMOTE, Borderline SMOTE, SMOTE-ENN, and CTGAN were used.

Preprocessing was performed and split the dataset into training and testing subsets using "train_test_split". To make sure it can be reproduced the parameter "random_state" was used with an 80% training and 20% test split. This evaluates the models as robust and reliable.

3.3 Synthetic Minority Oversampling Technique (SMOTE)

One data augmentation technique for class imbalance is SMOTE, which stands for Synthetic Minority Oversampling Technique and is meant to create synthetic examples of the minority class [14]. Furthermore, it is very effective when it comes to datasets in which the scale of one of the classes is substantially greater than that of the other. SMOTE is different from simple duplication techniques like doubling data, which can lead to overfitting, as new instances are created by interpolation between minority class samples. In addition to balancing the dataset, this approach helps preserve the underlying distribution of the minority class and improves the generalizability of a classifier on the data.

SMOTE draws synthetic data points for the minority class based on the neighboring samples. In this process, the model learns better during training thanks to better representation of the minority class and this results in better generalization. The method executes by taking N as an integer, which signifies the degree of oversampling that should be done to produce a balanced dataset with an equal number of instances per class. This parameter makes

it possible to represent a minority class while retaining the dataset's structure. The methodology involves three sequential and iterative steps, outlined below:

- A random sample is chosen from the minority class, which has fewer instances compared to the majority class.
- The k-nearest neighbors of the selected sample are identified, where k is a predefined number of closest data points.
- N neighbors are randomly selected from the k-nearest neighbors, and new synthetic samples are created by interpolating between the original sample and its chosen neighbors.

3.4 Borderline-SMOTE

There exists one advanced oversampling method called borderline-SMOTE [32] for the problem of class imbalance in datasets. Different from other methods, Borderline-SMOTE considers generating synthetic samples only for the minority class and priority is done on the borderline instance, i.e., instances lying in proximity to the decision boundary between two classes. The technique improves the definition of the decision boundary in these critical regions by creating new samples, which helps classifiers to identify cancerous from non-cancerous cases. This approach not only better represents the minority class, but also strengthens the generalization, resulting in more accurate model predictions.

3.5 SMOTE-ENN

This variant of SMOTE introduced by Batista and Prati is a hybrid sampling technique which is a combination of SMOTE and Edited Nearest Neighbor (ENN) [18]. The application of SMOTE to synthesize samples and ENN for removing noisy data points makes this method efficiently tackle the problem of imbalanced datasets. Especially, ENN removes data instances whose class labels disagree with the majority vote of their nearest neighbors and this improves data quality. SMOTE-ENN takes a mixture of these approaches to minimize the possibility of overfitting since the synthetic samples used for the minority class will be screened and therefore not allow the sample to overlap with the majority class space. The distinctive feature of this hybrid strategy is a more balanced and refined data set, which overall improves the performance and reliability of the model.

3.6 Conditional Tabular GAN (CTGAN)

Conditional Tabular Generative Adversarial Network (CTGAN) [33] is one of the pioneering works in the field of deep learning. This model has general applications on several data types such as images, audio, and text, making it a central point for future deep learning research. The two core components of CTGAN are the generator and the discriminator. The discriminator labels authentic and synthetic data and tries to distinguish their differences, while the generator fabricates synthetic data that looks realistic based on the real dataset. In the future, GANs will face challenges in modeling structured or tabular data, which are often non-Gaussian and even multimodal. CTGAN addresses this challenge through mode-specific normalization techniques.

A major problem in datasets is that the minority classes are highly underrepresented compared to the majority classes. The imbalance in such datasets can be partially compensated with conditional GANs that feature a conditional generator. The conditional generator forces the synthetic samples to adhere to the target classes or categories to balance the dataset and improve the model's ability to learn from previously underrepresented classes. This innovation improves the robustness and applicability of GANs to the imbalanced dataset scenario.

To ensure the robustness of conditional generation within the Conditional Generative Adversarial Network (CGAN) framework, three critical challenges must be addressed effectively.

- The condition must be effectively represented and seamlessly integrated into the generator's input to guide the synthetic data generation process.
- The generated samples must precisely align with the specified category or condition to ensure their relevance and usefulness.

The CTGAN combines the strengths of CGAN and TGAN and provides great power to cope with class imbalance and complex distributions of structured data. Using conditional generation, CTGAN has complete control over the class label assigned to synthetic samples, solving the class imbalance issue generally found in data sets. Furthermore, the network architecture design of CTGAN leads to an overall better model. It manages to accurately capture and replicate complex patterns in tabular datasets, and generate high-quality synthetic data. This capability makes CTGAN a valuable tool for various applications in data science and machine learning research.

3.7 ML Models

In this study, different ML models such as XGBoost, ETC, SVM, KNN, SGDC, LR, DT, RF, and GBC are explored for their use in the early diagnosis of lung cancer. The hyperparameters of each model were carefully optimized to improve the performance. Various hyperparameter configurations were utilized for these hyperparameters, and specific configurations were put in place to achieve the desirable accuracy and efficiency in the classification process to ensure that it is a reliable and effective process.

LR is a supervised learning algorithm used to predict categorical outcomes [34]. It is highly suitable for classification tasks and predicting analysis on large datasets by calculating the probability of class membership based on input variables. When the target variable is binary (i.e., it has two possible outcomes), this method is quite effective. Logistic regression is powerful and easy to understand despite being a simple approach, and its use in classification problems has been widespread.

The class of a test sample in KNN is determined by finding the closest data points (i.e. the nearest neighbors) and evaluating whether these k neighbors (i.e. the neighborhood] comprise points of class 1 or class 2 [35]. The basic concept of KNN is that a point should be classified according to the majority of its nearest neighbor in the feature space. The algorithm finds the k-nearest neighbors training points to the test sample using measuring the distance like Euclidean distance or some other similarity measure. The algorithm assigns the test point to the class that appears most frequently among the k neighbors. KNN is a very simple yet versatile algorithm, as it is a non-parametric technique that does not make assumptions about the underlying data distribution. Due to this flexibility, it is suitable for a wide variety of classification problems.

XGBoost is a supervised learning algorithm that is very efficient in handling regression and classification problems. Xgboost is known for its computational efficiency and is one of the favorite algorithms among data scientists [36]. In essence, XGBoost is a gradient-boosting method that creates an ensemble of decision trees. In the end, multiple trees' output is aggregated together, and each tree provides a score for the final prediction. This method makes XGBoost a more accurate and powerful choice for predictive modeling.

NB classifier works based on feature independence and classifies using the probabilistic approach. To enrich your training data, the assumption of independence does not hold at most of the real-world datasets [37] yet it is still heavily used and proves to be significantly useful most of the time. Bayes' Theorem can be used to calculate prior and likelihood probabilities, as well as posterior probabilities. Given the observed features, using these posterior probabilities, it predicts the most likely class. NB is highly regarded because despite the simplifying assumptions it makes, it is very reliable and efficient, and hence popular in classification tasks.

SVM is a very flexible method that is often used for data modeling and classification tasks. These can handle binary as well as multi-class classification problems by making use of kernel functions [38]. In the SVM algorithm, the main steps are the preprocessing of training data to normalize features and remove noise, next an appropriate kernel is selected. By maximizing the margin and identifying support vectors, the kernel transforms the data down into feature space and is able to separate classes with fewer errors. The parameter C is key to this trade-off between

maximizing margin and minimizing error on classification. When predicting, SVM classifies a new data point into a class by comparing its side of the hyperplane to the one calculated in training. If the data point falls on one side, the system classifies it into that class, and if it falls on the other side, it receives a different classification.

A widely used classification technique is DT which can be visualized (furthermore) as a tree-like structure, each node of which refers to a decision or prediction [39]. Recursively splitting the data in each node based on the most informative input features is what the algorithm does. This process breaks down the dataset into smaller subsets and branches into more refined classifications. The Gini Index and entropy are used as impurity measures at every node to construct the tree by evaluating the homogeneity of the data. The goal is to select relevant features to induce a tree that produces accurate and understandable predictions. It allows decision trees to be a powerful and intuitive tool for classification tasks.

For binary classification tasks, AdaBoost stands as one of the important ensemble learning algorithms documented in "adaboost" [40]. The system combines different weak learning algorithms to produce a strong final predictive model. To create each successive model AdaBoost assesses how previous models performed and selects misclassified cases to adjust the coming model for better error correction. The operation repeats itself until it generates the specified number of models or reaches the target performance level. The main strength of AdaBoost produces superior predictions from the combination of its component base models. The distinctive feature of AdaBoost as opposed to other ensemble techniques and machine learning approaches is its strong resistance to overfitting. Decision trees, which have a single splitting operation, serve as base learners for this algorithm. Similar to XGBoost, AdaBoost uses weights to boost the importance of misclassifications in successive iterations with updated weighting systems for prediction improvements. The classification process stops when the classification error reaches its minimum threshold or when the specified iteration limit is reached.

RF operates as an ensemble learning algorithm dedicated for binary classification according to [41]. RF constructs accurate predictors through its process of combining multiple weak learner base models. The algorithm conducts a special iteration to find and fix classification errors made by previous models thus enabling the succeeding models to address these identified errors. The method continues until it fulfills both design requirements, either by achieving a defined model threshold or meeting specific accuracy criteria. RF produces better predictions than its constituent parts through the process of combining all models so their collective forecasts produce an improved final model. This method functions effectively as a popular technique that practitioners use regularly in ML applications.

SGD serves as one of the prime optimization methods which iteratively modifies model parameters to decrease the cost function [42]. One significant variant of Gradient Descent introduces randomness through SGD because it utilizes one training sample at each iteration. The stochastic algorithm enables faster and more effective parameter adjustments through numerous updates than the Gradient Descent's approach to process a complete dataset per iteration. Focusing on just one training sample brings beneficial results to the model. As a result of processing data samples one by one SGD manages high computational efficiency which benefits large datasets and numerous parameter models. The algorithm starts by moving toward the local cost function minimum which it then continues to update using individual data points. The practicality and scalability of SGD enable its use for optimizing complex models in various applications.

ETC operates as a popular ensemble learning technique that uses multiple DTs to generate heightened accuracy levels according to [43]. A variant to the RF method where each split uses randomly selected important features to lower dependence on one feature and eliminate noise. The algorithm uses the Gini index as its detection method to find optimal features to split data before assigning Gini scores for the ranking of the importance of the characteristics. The method selects trees from a wide range of possibilities that increase model reliability and variety through diverse tree analysis.

GBC functions as an ML algorithm that combines various weak learners to build advanced predictive models according to [44]. GBC follows a proven principle which states that model aggregation produces superior predictive outcomes than using a standalone model. The predictive framework of GBC unites numerous weak learners which most commonly use decision trees to create a robust prediction model. GBC builds predictive models through

successive steps that optimize previous models' prediction errors until reaching optimal accuracy. The recursive refinement process transforms GBC into an exceptional method for analyzing complex classifying problems.

3.8 Evaluation Parameters

Any experiment-driven research is critically depend on assessing the performance of learning models. When applied to binary classification tasks, the popular evaluation metrics include precision, F1-score, recall, and accuracy. The following are the mathematical expressions for the metrics mentioned above.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{1}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

True Positives (TP) is defined as the cases when the model predicts the positive outcome correctly like predicting a disease in an affected person. The False Negative (FN) represents incorrect negative assessments of positive outcomes while the False Positive (FP) signifies incorrect positive predictions by the model for instance when it diagnoses healthy people with disease. False Negatives (TP) are cases whereby the model correctly predicts a negative outcome, for example, predicting that a person does not have a specific condition. False Negatives (FN) arise when the model fails to predict a positive outcome, meaning it fails to identify the presence of a disease in an infected individual. Metrics such as this are vital in determining the performance of a model, especially in critical domains, for example, healthcare, where an accurate prediction is of vital importance.

4 Results with Discussion

4.1 Experimental Configuration

To experiment within the context of detecting lung cancer, the data was split into 70% of the data in a training set and 30% of the data in a test set. To prevent overfitting and achieve a robust model performance, a standardized distribution strategy was used. To be able to verify the model's efficiency key metrics, like accuracy, precision, recall, and F1 score, were utilized. These experiments were performed in a Python environment with the use of different libraries, and executed on a high-performance Dell PowerEdge T430 server with a GPU. To support this, dual Intel Xeon processors with 8 cores at 2.4 GHz, and 32 GB DDR4 memory, combined with a GPU that featured 2 GB RAM were present in the system. The availability of these computational resources made the analytical and validation analysis processes fast and reliable, and in turn, made it possible to create reliable and accurate models of lung cancer diagnosis.

Table 3 Performance outcomes of ML models applied to the full dataset

Classifier	Accuracy	Precision	Recall	F1-Score
XGB	0.6759	0.68	0.68	0.67
ETC	0.7962	0.80	0.80	0.80
SVM	0.5648	0.56	0.56	0.56
KNN	0.6759	0.68	0.68	0.67
SGDC	0.6759	0.68	0.68	0.68
LR	0.6944	0.69	0.69	0.69
DT	0.6759	0.68	0.68	0.67
RF	0.7407	0.74	0.74	0.74
GBC	0.6851	0.69	0.69	0.68
NB	0.6666	0.67	0.67	0.67

4.2 ML Model Results

4.2.1 ML Models Results with Complete Dataset

This section presents the results of applying various ML models to the comprehensive lung cancer dataset. A detailed breakdown of the results is provided in Table 3.

Different ML models demonstrated significantly different results when used to detect lung cancer. The SVM provided the least accurate predictions at 0.5648 alongside 0.56 precision rate and recall values and F1-score. The NB showed modest advancement by reaching a 0.6666 accuracy score. XGBoost, KNN, and SGDC together with DT showed similar performance levels where accuracy reached 0.6759 and precision and recall scored 0.68 while F1-score scored 0.67. The GBC provided a slight accuracy gain by reaching 0.6851. LR reached the highest performance with 0.6944 accuracies and 0.69 scores for precision, recall, and F1-score. The RF model displayed improved performance by reaching 0.7407 accuracy and a precision value of 0.74. The ETC demonstrated the best performance among the studied models in lung cancer detection by achieving an accuracy rate of 0.7962 and 0.80 values of precision, recall, and F1-score.

4.2.2 Performance Comparison of ML Models with SMOTE

The study utilized different upsampling approaches to handle the significant class discrepancy present in the research dataset because unbalanced data produces skewed results. The SMOTE was used to equalize sample distribution in the dataset. The dataset acquired 270 cancer case samples and 270 no-cancer instances after SMOTE processing thus achieving equal class distributions. The performance results using the SMOTE-balanced dataset for ML models are presented in Table 4.

SMOTE-based class imbalance treatment improves various ML models for diagnosing lung cancer by enhancing their performance. The SVM demonstrated an accuracy of 0.6574 along with a precision of 0.69 and recall of 0.66 and an F1-score of 0.65. SGDC achieved an accuracy of 0.6666 and precision of 0.78 alongside a recall of 0.67 which resulted in a 0.64 F1-score. According to the results, the KNN model showed a significant advancement reaching 0.9166 accuracy alongside 0.92 precision and F1-score and recall. The NB model showed outstanding performance with 0.9259 accuracy and 0.93 value of precision, recall, and F1-score. The DT model reached an outstanding accuracy score of 0.9444 and 0.94 scores of precision, recall, and F1 scores. The XGBoost, ETC, LR, RF, and GBC performed identically by attaining an accuracy of 0.9537. Their precision, F1-score, and recall values remained constant at 0.95. The accuracy results for lung cancer detection became significantly enhanced through SMOTE implementation while multiple models demonstrated outstanding results that approached perfect accuracy.

 https://doi.org/10.1007/s44196-025-00879-4

Table 4 The results of ML models with SMOTE

Classifier	Accuracy	Precision	Recall	F1-Score
XGB	0.9537	0.95	0.95	0.95
ETC	0.9537	0.95	0.95	0.95
SVM	0.6574	0.69	0.66	0.65
KNN	0.9166	0.92	0.92	0.92
SGDC	0.6666	0.78	0.67	0.64
LR	0.9537	0.95	0.95	0.95
DT	0.9444	0.94	0.94	0.94
RF	0.9537	0.95	0.95	0.95
GBC	0.9537	0.95	0.95	0.95
NB	0.9259	0.93	0.93	0.93

Table 5 The results of ML models with borderline SMOTE upsampling

Classifier	Accuracy	Precision	Recall	F1-Score
XGB	0.9629	0.97	0.96	0.96
ETC	0.9537	0.95	0.95	0.95
SVM	0.6018	0.63	0.60	0.59
KNN	0.9351	0.94	0.94	0.94
SGDC	0.6111	0.76	0.61	0.56
LR	0.9444	0.94	0.94	0.94
DT	0.9444	0.95	0.94	0.94
RF	0.9537	0.95	0.95	0.95
GBC	0.9351	0.94	0.94	0.94
NB	0.9351	0.94	0.94	0.94

4.2.3 ML Models Results Using Borderline SMOTE

Borderline SMOTE served as the advanced version of SMOTE that we used to resolve class imbalance during these experiments. Borderline SMOTE application produced a dataset containing 540 instances in which 270 belonged to the cancer class and 270 to the non-cancer class thus achieving data balance. The experimental results of ML model performance are shown in Table 5.

The application of the Borderline SMOTE technique improves different levels of performance in ML models designed for lung cancer detection. The SVM model operating at its lower performance level reaches 0.6018 accuracy alongside 0.63 precision, 0.60 recall, and 0.59 F1-score. The accuracy score of SGDC reaches 0.6111, precision at 0.76, recall at 0.61, and F1-score at 0.56. The KNN, NB, and GBC achieve similar performance in the mid-range with all models reaching 0.9351 accuracy and maintaining 0.94 for precision, recall, and F1-scores. The LR model shares accuracy statistics of 0.9444 with DT where LR displays precision, recall, and F1-score at 0.94 and DT reveals higher precision at 0.95 compared to LR. The performance numbers for RF and ETC match since both models achieve 0.9537 accuracy alongside 0.95 precision, recall, and F1-score. XGBoost reaches the highest accuracy rate of 0.9629 and proves to be the most effective model. Borderline SMOTE boosts model performance for most designs and XGBoost emerges as the best method for lung cancer diagnosis.

4.2.4 ML Models Results Using SMOTE-ENN

The experimental phase involved utilizing the SMOTE-ENN technique for handling class imbalance issues. The implementation of SMOTE-ENN achieved dataset balancing which transformed the total instances into 466 samples. Of the total 466 instances analyzed, 223 belonged to the cancer category and the non-cancer category consisted

Table 6 The results of ML models with SMOTE-ENN upsampling

Classifier	Accuracy	Precision	Recall	F1-Score
XGB	0.9838	0.98	0.98	0.98
ETC	0.9677	0.97	0.97	0.97
SVM	0.9677	0.94	0.97	0.95
KNN	0.9516	0.96	0.95	0.96
SGDC	0.9838	0.98	0.98	0.98
LR	0.9677	0.97	0.97	0.97
DT	0.9677	0.97	0.97	0.97
RF	0.9677	0.97	0.97	0.97
GBC	0.9516	0.96	0.95	0.96
NB	0.9516	0.96	0.95	0.96

Table 7 The results of ML models with CTGAN upsampling

Classifier	Accuracy	Precision	Recall	F1-score
XGB	0.9787	0.98	0.98	0.98
ETC	0.9787	0.98	0.98	0.98
SVM	0.6914	0.81	0.69	0.67
KNN	0.9574	0.96	0.96	0.96
SGDC	0.6489	0.77	0.65	0.62
LR	0.9787	0.98	0.98	0.98
DT	0.9787	0.98	0.98	0.98
RF	0.9893	0.99	0.99	0.99
GBC	0.9787	0.98	0.98	0.98
NB	0.9468	0.95	0.95	0.95

of 243 samples. This paper shows the performance results of machine learning models that use the dataset from Table 6.

All ML models employing SMOTE-ENN demonstrate exceptional accuracy results for detecting lung cancer. All three models including KNN, GBC, and NB demonstrate similar 0.9516 accuracy during the lower performance range. The precision value, F1-score, and recall metrics from KNN and GBC measurement reached 0.96 at the same time but NB exhibited precision at 0.96 along with recall at 0.95 while maintaining an F1-score at 0.96. SVM along with LR, DT, RF, and ETC achieves a 0.9677 accuracy level. The evaluation metrics for SVM differ from the metrics used by LR, DT, RF, and ETC because all four models display precision at 0.97 but they share an equal F1-score and recall at 0.97. XGB and SGDC demonstrate superior model performance in this task by achieving 0.9838 accuracy accompanied by complete precision and recall values. These results make it clear that implementing SMOTE-ENN improved performance, as XGBoost and SGDC were found to be the best models as measured by all evaluation criteria.

4.2.5 ML Models Results Using CTGAN

The last step of the experiment employed CTGAN to solve problems caused by class imbalance. The dataset obtained 540 instances following the CTGAN application with equal distributions of 270 cancer instances and 270 non-cancer instances. The ML model performance using CTGAN-balanced data appears in Table 7.

Different accuracy levels result in varying effectiveness of ML models that detect lung cancer with CTGAN. The SGDC reaches 0.6489 accuracy while producing a precision of 0.77 and recall of 0.65 along with an F1-score of 0.62 at its performance level. The SVM attains accuracy rates of 0.6914 but also reaches precision at 0.81 and recall at 0.69 and finally establishes an F1-score of 0.67. The KNN model reaches precision, recall, and F1-score values of 0.96 while maintaining a 0.9574 accuracy level. The NB model shows identical accuracy results with

 https://doi.org/10.1007/s44196-025-00879-4

Table 8 Accuracy comparison across classifiers

Classifier	CTGAN	SMOTE-ENN	Borderline SMOTE	SMOTE	Original
XGB	0.9787	0.9838	0.9629	0.9537	0.6759
ETC	0.9787	0.9677	0.9537	0.9537	0.7962
SVM	0.6914	0.9677	0.6018	0.6574	0.5648
KNN	0.9574	0.9516	0.9351	0.9166	0.6759
SGDC	0.6489	0.9838	0.6111	0.6666	0.6759
LR	0.9787	0.9677	0.9444	0.9537	0.6944
DT	0.9787	0.9677	0.9444	0.9444	0.6759
RF	0.9893	0.9677	0.9537	0.9537	0.7407
GBC	0.9787	0.9516	0.9351	0.9537	0.6851
NB	0.9468	0.9516	0.9351	0.9259	0.6666

KNN since it reaches an accuracy rate of 0.9468. The combination of XGBoost with ETC, LR, DT, and GBC achieves an accuracy rate of 0.9787 alongside a precision score of 0.98 and a recall score of 0.98 resulting in equally matched F1-score values. The RF model exhibited the best performance with a 0.9893 accuracy level and complete precision, recall, and F1-score measurements at 0.99. CTGAN demonstrates remarkable capabilities for model detection enhancement through its application which makes RF emerge as the optimal choice across all evaluation metrics.

CTGAN outperforms SMOTE-ENN due to its ability to generate more realistic and diverse synthetic data by learning the underlying joint probability distribution of real data through GAN-based deep learning. Unlike SMOTE-ENN, which uses linear interpolation and may struggle with categorical features, CTGAN handles both continuous and categorical variables effectively, making it more suited for complex datasets like those in medical diagnostics. Additionally, CTGAN avoids information loss by preserving subtle patterns in the data while being more robust to noise and outliers. This diverse sample generation mitigates overfitting and enhances classifier generalization. As a result, CTGAN improves classification metrics, especially in imbalanced datasets such as those used for lung cancer detection.

4.3 Comparison of Results

A comprehensive performance evaluation of the models required the analysis of accuracy from all sets of experiments. Different upsampling approaches lead to substantial enhancements in accuracy levels for ML models as shown through direct analysis. Table 8 shows how different data re-sampling methods affect the accuracy performance of ML models through direct comparisons.

SVM demonstrated the lowest accuracy rate at 0.5648 when testing the original dataset because the other models including SGDC scored 0.6759 and XGB and KNN achieved equal scores at 0.6759. The performance levels increased after implementing SMOTE because SGDC achieves accuracy at 0.6666 The accuracy from SVM reached 0.6574 while NB produced accuracy at 0.6666. Model accuracy increases to 0.6018 after Borderline SMOTE improves SVM performances and reaches 0.6111 accuracy for SGDC with 0.9351 accuracy for NB. The combination of SMOTE-ENN methodology enabled XGB and SGDC to match the 0.9838 accuracy rate which remained equal to SVM's 0.9677 value of accuracy. With CTGAN the models can generate their best results as SGDC reaches 0.6489 accuracy alongside SVM achieving 0.6914 accuracy and RF reaching 0.9893. Every experiment shows RF as the leading model while both CTGAN and SMOTE-ENN enhance model performance.

4.4 Significance of Proposed Work

Early detection of lung cancer is critical because it improves the likelihood of diagnosing the disease at an early stage, when it is most treatable, and could improve the survival rates. Recent research works state that lung cancer

Table 9 Results of K-fold cross-validation

Folds	Accuracy	Precision	Recall	F1-score
First-fold	0.9868	0.9898	0.9893	0.9867
Second-fold	0.9896	0.9845	0.9828	0.9835
Third-fold	0.9859	0.9848	0.9848	0.9848
Fourth-fold	0.9878	0.9835	0.9666	0.9759
Fifth-fold	0.9896	0.9828	0.9666	0.9848
Average	**0.9879**	**0.9851**	**0.9780**	**0.9831**

The bold values indicating average of all above 5 fold results, indicating significance of proposed model

patient survival at Stage 1 is much higher than Stage 2. The AI-driven methods, such as the proposed CTGAN-RF framework, play a transformative role in early diagnosis by accurately identifying patterns in clinical data that may be ignored in traditional diagnostic procedures. By addressing class imbalance issues in medical datasets using synthetic data augmentation, the proposed framework minimizes false negatives, thereby reducing the risk of missed diagnoses. Following this framework, high-risk patients' survival is increased with timely interventions. With early detection, accurate treatment strategies can be developed for patient survival. This alignment with precision medicine improves therapeutic outcomes, reduces side effects, and enhances overall patient management.

The integration of CTGAN-generated synthetic data significantly improves predictive performance because CTGANs have a more complex structure for better understanding, non-linear relationships, and high-quality synthetic data. Unlike SMOTE or standard resampling methods, which generate synthetic samples using minority-class data with a linear interpolation technique. The CTGANs provide better realistic data by focusing on outlier patterns and correlations. Furthermore, deep learning models excel in feature extraction for large, balanced datasets. These datasets require extensive computational resources and struggle with small, imbalanced datasets. Results reveal that the CTGAN-RF framework outperforms deep learning models on the lung cancer dataset in terms of accuracy, precision, recall, and F1-score, particularly due to RF's robustness with tabular features and the effective synthetic data generation by CTGAN. Additionally, RF has lower memory consumption, making it more efficient and accessible for real-time deployment in resource-constrained environments, such as mobile health applications and edge devices.

4.5 K-Fold Cross-Validation Results

This research uses Table 9 to present results from the k-fold cross-validation evaluation of the RF model. The evaluation assesses the performance of the RF model by applying it to previously unseen data. The newly generated feature set went through validation testing with five separate splits for the classifier. The RF techniques produced a k-fold accuracy rate of 0.9879 which demonstrated impressive results. Our RF model, which incorporated CTGAN, demonstrated the best accuracy levels of 0.9893. The research demonstrates the robust validation capabilities and generalization performance of the implemented techniques which prove effective in identifying lung cancer cases.

4.6 Comparison with SOTA

Table 10 demonstrates an elaborate comparison between our research approach and existing studies in publications. An analysis of publications within the 2022 to 2024 timeframe received priority consideration to achieve fairness in research. Researchers discovered that the highest reported performance using ML techniques at 0.98 was not sufficient based on their evaluation. A new ML algorithm incorporated CTGAN technology to boost its functionality in our research. The combination of RF with CTGAN upsampling achieved a 0.9893 accuracy rate in lung cancer diagnosis. The proposed innovation surpasses current approaches based on the comparative evaluation results.

Table 10 Performance comparison with state-of-the-art techniques

References	Year	Dataset	Techniques	Accuracy
[21]	2024	Bangla dash dataset	XGBoost	0.9692
[22]	2022	Kaggle	RotF	0.971
[23]	2022	UCI	SVM with SMOTE	0.988
[24]	2023	Self-collected	RF	0.9032
[25]	2023	Kaggle	SVM	0.98
[28]	2022	Data world	SVM	0.96
[29]	2022	UCI	KNN with PCA features	0.87
[30]	2023	Kaggle	SVM	0.91
Proposed	**2024**	**Kaggle 309 instances**	**RF with CTGAN**	**0.9893**

The bold values indicates results of proposed model and shows the proposed model results are better than all SOTA techniques

4.7 Generalizability of Proposed Framework

To check the generalizability and stability of the proposed framework on other small-sized, imbalanced, and medical domain datasets, we tested the proposed framework on 3 other independent benchmark datasets. The dataset is named "Pima Indians Diabetes Dataset", having only 768 instances, in which normal patient records are more than diabetic [45]. The second dataset is "Breast Cancer Wisconsin (Diagnostic)" which contains 569 instances having fewer malignant cases as compared to benign [46]. The last dataset is "Heart Disease" with 303 instances, with fewer heart disease patient records than normal [47]. The proposed CTGAN-RF framework performs well on all three datasets by giving an accuracy score of 98.78% on the Pima Indians Diabetes Dataset, 99.92% on Breast Cancer Wisconsin (Diagnostic), and 99.99% on "Heart Disease". This stability and good accuracy show promising results of the proposed framework on all small-sized and imbalanced medical domain datasets.

5 Conclusion

Experimental data from this research demonstrates that a combination of CTGAN-generated features with Random Forest classification holds major potential for lung cancer detection. The model achieved an exceptional performance level because it produced an accuracy score of 0.9893 alongside precision, recall and F1 scores that each measured at 0.99. The model outperforms standard machine learning classifiers SVM, k-NN, and DT by producing exceptional results mainly when supported by data balancing methods SMOTE and Borderline-SMOTE and SMOTE-ENN. The combination of the CTGAN-Random Forest model with its capabilities to manage class imbalance reveals synthetic data augmentation as a powerful solution. Cross-validation using five partitions together with assessments against competing methods showed this approach achieved reliable outcomes consistently. This research confirms how advanced data generation approaches combined with strong machine learning models support better lung cancer detection and diagnosis at early stages. Letting our approach deliver high marksmanship figures allows the development of dependable diagnostic solutions. The innovative approach leads to better patient results through quick and accurate detection methods which help doctors create individualized powerful treatment plans. Research should explore these techniques on more medical data types and apply them to other diagnosis challenges, which would push medical diagnosis research forward.

Author Contributions NI conceptualization, formal analysis, writing—the original manuscript. AA data curation, conceptualization, writing—the original manuscript. DAA writing—review and editing, formal analysis, and funding. AH software, methodology, and project administration. MU investigation, software, and visualization. SA visualization, resources, investi-

gation. ST methodology, funding, validation, writing—review, and editing. IA supervision, validation, writing—review, and editing. All authors reviewed the article and approved it.

Funding Nisreen Innab would like to express sincere gratitude to AlMaarefa University, Riyadh, Saudi Arabia, for supporting this research. The authors are thankful to Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R508), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/379/46.

Data Availability The dataset can be requested from the corresponding authors.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no conflict of interest.

References

1. World Health Organization.: Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer
2. Tomassini, S., Falcionelli, N., Sernani, P., Burattini, L., Dragoni, A.F.: Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: A survey. Comput. Biol. Med. **146**, 105691 (2022). https://doi.org/10.1016/j.compbiomed.2022.105691
3. Guo, M., et al.: Autologous tumor cell-derived microparticle-based targeted chemotherapy in lung cancer patients with malignant pleural effusion. Sci. Transl. Med. (2019). https://doi.org/10.1126/scitranslmed.aat5690
4. Ahamed, M.F., Shafi, F.B., Nahiduzzaman, M., Ayari, M.A., Khandakar, A.: Interpretable deep learning architecture for gastrointestinal disease detection: a tri-stage approach with PCA and XAI. Comput. Biol. Med. **185**, 10950 (2025). https://doi.org/10.1016/j.compbiomed.2024.109503
5. Ahamed, M.F., Nahiduzzaman, M., Islam, M.R., Naznine, M., Arselene Ayari, M., Khandakar, A., et al.: Detection of various gastrointestinal tract diseases through a deep learning method with ensemble ELM and explainable AI. Expert Syst. Appl. **256**, 12490 (2024). https://doi.org/10.1016/j.eswa.2024.124908
6. Ahamed, M.F., Hossain, M.M., Nahiduzzaman, M., Islam, M.R., Islam, M.R., Ahsan, M., et al.: A review on brain tumor segmentation based on deep learning methods with federated learning techniques. Comput. Med. Imaging Graph. **110**, 10231 (2023). https://doi.org/10.1016/j.compmedimag.2023.102313
7. Ahamed, M.F., Salam, A., Nahiduzzaman, M., Abdullah-Al-Wadud, M., Islam, S.R.: Streamlining plant disease diagnosis with convolutional neural networks and edge devices. Neural Comput. Appl. **36**(29), 18445–18477 (2024)
8. Ahamed, M.F., Islam, M.R., Nahiduzzaman, M., Karim, M.J., Ayari, M.A., Khandakar, A.: Automated detection of colorectal polyp utilizing deep learning methods with explainable AI. IEEE Access **12**, 78074–7810 (2024). https://doi.org/10.1109/ACCESS.2024.3402818
9. Hossain, M.M., Islam, M.R., Ahamed, M.F., Ahsan, M., Haider, J.: A collaborative federated learning framework for lung and colon cancer classifications. Technologies (2024). https://doi.org/10.3390/technologies12090151

10. Sarkar, O., Islam, M.R., Syfullah, M.K., Islam, M.T., Ahamed, M.F., Ahsan, M., et al.: Multi-scale CNN: an explainable AI-integrated unique deep learning framework for lung-affected disease classification. Technologies (2023). https://doi.org/10.3390/technologies11050134
11. Ahamed, M.F., Syfullah, M.K., Sarkar, O., Islam, M.T., Nahiduzzaman, M., Islam, M.R., et al.: IRv2-Net: a deep learning framework for enhanced polyp segmentation performance integrating InceptionResNetV2 and UNet architecture with test time augmentation techniques. Sensors (2023). https://doi.org/10.3390/s23187724
12. Wang, Y., et al.: Gene selection from microarray data for cancer classification—a machine learning approach. Comput. Biol. Chem. **29**(1), 4–37 (2005). https://doi.org/10.1016/j.compbiolchem.2004.11.001
13. Mandal, S., Banerjee, I.: Cancer classification using neural network. Int. J. **172**, 18–49 (2015)
14. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–35 (2002). https://doi.org/10.1613/jair.953
15. Li, W., et al.: EID-GAN: generative adversarial nets for extremely imbalanced data augmentation. IEEE Trans. Ind. Inform. **19**(6), 3208–321 (2022). https://doi.org/10.1109/TII.2022.3153188
16. Zieba, M., Tomczak, J.M.: Boosted SVM with active learning strategy for imbalanced data. Soft Comput. **19**, 336–3357 (2014). https://doi.org/10.1007/s00500-014-1370-3
17. He, H., Zhang, W., Zhang, S.: A novel ensemble method for credit scoring: adaption of different imbalance ratios. Expert Syst. Appl. **98**, 105–11 (2018). https://doi.org/10.1016/j.eswa.2017.10.023
18. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. **6**(1), 20–2 (2004). https://doi.org/10.1145/1007730.1007735
19. Li, D.C., Wang, S.Y., Huang, K.C., Tsai, T.I.: Learning class-imbalanced data with region-impurity synthetic minority oversampling technique. Inf. Sci. **607**, 140–1391 (2022). https://doi.org/10.1016/j.ins.2021.10.040
20. Fernandez, A., Garcia, S., Herrera, F., Chawla, N.V.: SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. **61**, 863–90 (2018). https://doi.org/10.1613/jair.1.11215
21. Bhuiyan, M.S., Chowdhury, I.K., Haider, M., Jisan, A.H., Jewel, R.M., Shahid, R., et al.: Advancements in early detection of lung cancer in public health: a comprehensive study utilizing machine learning algorithms and predictive models. J. Comput. Sci. Technol. Stud. **6**(1), 113–121 (2024)
22. Dritsas, E., Trigka, M.: Lung cancer risk prediction with machine learning models. Big Data Cogn. Comput. **6**(4), 139 (2022)
23. Anil Kumar, C., Harish, S., Ravi, P., Svn, M., Kumar, B.P., Mohanavel, V., et al.: Lung cancer prediction from text datasets using machine learning. BioMed Res. Int. **2022**(1), 6254177 (2022)
24. Mohan, K., Thayyil, B.: Machine learning techniques for lung cancer risk prediction using text dataset. Int. J. Data Inform. Intell. Comput. **2**(3), 47–56 (2023)
25. Fatoki, F.M., Akinyemi, E.K., Phlips, S.A.: Prediction of lungs cancer diseases datasets using machine learning algorithms. Curr. J. Appl. Sci. Technol. **42**(11), 15–23 (2023)
26. Wayahdi, M.R., Ruziq, F.: KNN and XGBoost algorithms for lung cancer prediction. J. Sci. Technol. (JoSTec) **4**(1), 179–186 (2022)
27. Nabeel, S.M., Bazai, S.U., Alasbali, N., Liu, Y., Ghafoor, M.I., Khan, R., et al.: Optimizing lung cancer classification through hyperparameter tuning. Digit. Health **10**, 20552076241249660 (2024)
28. Vasudha Rani, V., Das, S., Kundu, T.K.: Risk prediction model for lung cancer disease using machine learning techniques. In: Innovations in Computer Science and Engineering: Proceedings of the Ninth ICICSE, pp. 417–425. Springer, Singapore (2022)
29. Gültepe, Y.: Performance of lung cancer prediction methods using different classification algorithms. Comput. Mater. Continua **67**(2), 2015–2028 (2021). https://doi.org/10.32604/cmc.2021.014631
30. Dirik, M.: Machine learning-based lung cancer diagnosis. Turk. J. Eng. **7**(4), 322–330 (2023)
31. Hugging Face Datasets.: Lung Cancer Dataset. https://huggingface.co/datasets/nateraw/lung-cancer. Accessed 16 May 2024
32. Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Springer (2005)
33. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32, pp. 7335–7345. Curran Associates Inc, San Francisco (2019)
34. Menard, S.: Applied Logistic Regression Analysis, vol. 106. Sage, Philadelphia (2002)
35. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
36. Chen, T., He, T., Benesty, M., Khotilovich, V., Cho, H., Tang, T., et al.: Xgboost: extreme gradient boosting. R package version 04-2, 1(4), 1–4 (2015)
37. Rish, I., et al.: An empirical study of the Naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, pp. 41–46 (2001)
38. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intell. Syst. Appl. **13**(4), 18–28 (1998)

39. Song, Y.Y., Ying, L.: Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry **27**(2), 130 (2015)
40. Hastie, T., Rosset, S., Zhu, J., Zou, H.: Multi-class adaboost. Stat. Interface **2**(3), 349–360 (2009)
41. Belgiu, M., Drăguţ, L.: Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogramm. Remote Sens. **114**, 24–31 (2016)
42. Taha, A.A., Malebary, S.J.: An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access **8**, 25579–25587 (2020)
43. Umer, M., Sadiq, S., Nappi, M., Sana, M.U., Ashraf, I., et al.: ETCNN: extra tree and convolutional neural network-based ensemble model for COVID-19 tweets sentiment classification. Pattern Recognit. Lett. **164**, 224–231 (2022)
44. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Front. Neurorobot. **7**, 21 (2013)
45. Repository UML.: Pima Indians Diabetes Database. Contains numerical and categorical features (glucose, BMI, age) with 768 instances. Class imbalance present in positive vs. negative diabetes cases. https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes. Accessed on 10 Jan 2025
46. Repository UML, Kaggle.: Breast Cancer Wisconsin (Diagnostic) Dataset. Includes 569 instances with numerical (radius, texture, smoothness) and categorical features. Imbalanced class with fewer malignant cases. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic). Accessed on 10 Jan 2025
47. Repository UML.: Cleveland Heart Disease Dataset. Comprises 303 instances with numerical features (cholesterol, heart rate) and categorical features. Imbalanced positive vs. negative heart disease cases. https://archive.ics.uci.edu/ml/datasets/heart+disease. Accessed on 10 Jan 2025

Authors and Affiliations

Nisreen Innab[1] · Asma Aldrees[2] · Dina Abdulaziz AlHammadi[3] · Abeer Hakeem[4] · Muhammad Umer[5] · Shtwai Alsubai[6] · Silvia Trelova[7] · Imran Ashraf[8]

✉ Imran Ashraf
imranashraf@ynu.ac.kr

Nisreen Innab
Ninnab@um.edu.sa

Asma Aldrees
edrees@kku.edu.sa

Dina Abdulaziz AlHammadi
daalhammadi@pnu.edu.sa

Abeer Hakeem
Ahakim@kau.edu.sa

Muhammad Umer
umer.sabir@iub.edu.pk

Shtwai Alsubai
Sa.alsubai@psau.edu.sa

Silvia Trelova
silvia.trelova@fm.uniba.sk

1 Department of Computer Science and Information Systems, College of Applied Sciences, AlMaarefa University, Diriyah, 13713 Riyadh, Saudi Arabia

2 Department of Informatics and Computer Systems College of Computer Science, King Khalid University, Abha, Saudi Arabia

3 Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, 11671 Riyadh, Saudi Arabia

[4] Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

[5] Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

[6] Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, 11942 Al-Kharj, Saudi Arabia

[7] Faculty of Management, Comenius University Bratislava, Bratislava 25, 10, 82005 Odbojarov, Slovakia

[8] Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea