RESEARCH ARTICLE

Early Detection of Lung Cancer Using Predictive Modeling Incorporating CTGAN Features and Tree-Based Learning

ABDULRAHMAN ALZAHRANI

Department of Computer Science and Engineering, College of Computer Science and Engineering, University of Hafr Al Batin, Hafr Al Batin 39524, Saudi Arabia

e-mail: aalzahrani@uhb.edu.sa

ABSTRACT Artificial intelligence (AI) has revolutionized several domains and medical science has significantly benefited from AI-based approaches. Machine learning models enable the robust and accurate analysis of large streams of medical data, assisting medical experts in creating specialized and targeted treatment plans. Lung cancer also requires further automated detection tools as it remains one of the most lethal cancers worldwide, necessitating advancements in early and accurate detection methods. This study presents a novel approach to lung cancer classification by leveraging synthetic data generated using conditional tabular generative adversarial networks (CTGAN) and applying a random forest (RF) classifier. The proposed model has shown exceptional performance with an accuracy of 98.93%, and precision, F1 score, and recall of 99%. To validate the efficacy of the approach, we conducted extensive experiments using nine other classifiers, including support vector machines, k-nearest neighbors, decision trees, and others. These classifiers were evaluated with different data balancing techniques such as synthetic minority oversampling technique (SMOTE), borderline-SMOTE, and SMOTE-ENN, alongside original, unbalanced dataset. Comparative analysis revealed that the CTGAN-RF model outperforms these traditional classifiers, particularly in handling class imbalance and improving predictive accuracy. Additionally, the robustness of the proposed model was confirmed through 5-fold cross-validation, further emphasizing its reliability. The approach was benchmarked against state-of-the-art (SOTA) methods in lung cancer detection, highlighting significant improvements in classification metrics. This comprehensive evaluation highlights the potential of synthetic data augmentation combined with machine learning techniques in enhancing the early detection and diagnosis of lung cancer, opening the way for improved patient outcomes and personalized treatment strategies.

INDEX TERMS Predictive modeling, AI-drive disease detection, lung cancer classification, healthcare, bioinformatics, early disease detection.

I. INTRODUCTION

Artificial intelligence (AI) paradigm brought revolution in several domains including engineering, production, medical science, etc. Particularly, the use of machine learning approaches has greatly helped medical science develop automated solutions. Machine learning models help analyze large streams of medical data robustly and accurately helping

The associate editor coordinating the review of this manuscript and approving it for publication was Ayman El-Baz.

medical experts make specialized and targeted treatment plans. Among the most cancer-related deaths worldwide, lung cancer is among the leading ones. Diagnosis and treatment at the beginning stage significantly improve the survival rate of patients and reduce the mortality rate. In 2017, lung cancer accounted for the most significant number of disability-adjusted life years (DALYs), with 40.9 million patients suffering from this disease. It is estimated that as many as 1.8 million deaths occurred in 2020 [1], [2] due to lung cancer, and statistics show that as many as 135,720

patients would die from the disease in the United States itself; however, lung cancer does not initially show symptoms during its beginning and then gets quite complicated to be detected at an early stage. The current study focuses on analyzing and predicting the growth and progression of lung cancer using machine learning (ML) and deep learning (DL) approaches [3]. Around 10-16% of lung cancer occurs due to Small Cell Lung Cancer (SCLC). The SCLC development rate is very fast and ultimately results in tumors' rise, often metastasizing widely, which are developed in close association with cigarette smoking.

Machine learning (ML) algorithms can reliably predict whether a patient will develop lung cancer at an early stage by recognizing patterns of cancer and factors with the condition through the analysis of vast amounts of records of patients. Recent developments in bioinformatics (images and factor-based) have made early detection of lung cancer possible. The idea is to predict a patient's risk by extracting relevant features from these datasets and applying a robust ML algorithm. ML algorithms can significantly improve the accuracy of malignancy predictions. In [4], authors have applied several metrics like Confusion Matrix, Precision, F1 Score, Recall, and Accuracy, along with multiple algorithmic approaches such as Logistic Regression(LR), Random Forest(RF), Decision Tree(DT), Support Vector Machine(SVM), Naive Bayes(NB), and K-Nearest Neighbor(KNN). All these techniques are used to recognize gene expression signatures from samples and develop predictive cancer models.

The Authors designed and evaluated an ADB-based prognostic model to provide medical professionals with an ECOG PS-based decision-making process for lung cancer patients. Many scientific articles have addressed the application of ML in this particular disease. We propose a methodology for designing effective ML classification models for lung cancer prediction utilizing everyday routines and symptoms/signs as input characteristics for the models. AI and ML play significant role in modern healthcare. The efficacy of ML models is heavily reliant on how effectively the model can classify true and false positives and negatives based on the different datasets used for classification [5]. The study focuses on precision, accuracy, and F1 scores, considering a significant body of literature. The results of the current study discussed next, represent significant improvements compared with earlier studies [6], [7]. In particular, it will assess the efficiency of different classifiers like XGBoost, Extra Trees Classifier(ETC), SVM, KNN, Stochastic-Gradient-Descent-Classifier(SGDC), LR, DT, RF, and Gradient-Boosting Classifier(GBC) for early-detection of cases with lung cancer by analyzing the pertinent medical data [8].

The purpose of this review is to discuss research that uses multiple techniques for lung cancer prediction utilizing clinical risk factor data. It includes a summary of the survival prediction processes, feature selection and extraction techniques, and ML and DL algorithms with multiple validation metrics. Moreover, some researchers have proposed a handmade e-nose device for detecting lung cancer [9],

[10]. Despite all the work that has gone into this area, significant incentive still exists for further refinement of the classification methods. Neural networks have also been used to enhance classification performance, with feedforward and backpropagation methods being especially common in such attempts [11], [12].

A. RESEARCH GAPS

From the literature review, several gaps and limitations in previous studies were identified including the limited use of synthetic data generation for oversampling techniques like SMOTE, Borderline-SMOTE, and SMOTE ENN, which may not fully capture the distribution of minority class samples, leading to biased model performance. Most existing lung cancer prediction models struggle with the class imbalance present in medical datasets, leading to poor generalization in real-world scenarios. Few studies conducted extensive cross-validation to confirm the reliability of their models. Many models suffered from overfitting due to the lack of proper validation techniques. Thus, this study bridges these gaps by introducing a novel approach that integrates CTGAN-generated features with a robust RF classifier, improving predictive performance and addressing class imbalance.

B. RESEARCH AIMS

The primary aim of this study is to develop an improved predictive modelling approach for the early detection of lung cancer by leveraging CTGAN-generated synthetic data and tree-based learning techniques, particularly the Random Forest (RF) classifier. The objective is to address class imbalance challenges in existing datasets and enhance the robustness of machine learning-based detection models for lung cancer.

C. MAJOR CONTRIBUTIONS

The major contributions of this research work are:

- Developed novel framework for lung cancer disease detection using CTGAN-generated features in conjunction with a Random Forest classifier.
- Achieved outstanding performance metrics, with 98.93% accuracy, precision, F1 score, and recall of 99%, demonstrating significant improvements over traditional methods.
- Conducted comprehensive comparisons with nine traditional ML classifiers: XGB, ETC, SVM, KNN, SGDC, LR, DT, RF, and GBC, utilizing SMOTE, Borderline-SMOTE, SMOTE ENN, and original features for a robust analysis.
- Implemented extensive 5-fold cross-validation of the proposed model to ensure robustness and reliability compared with previously published research works.
- Highlighted the effectiveness of integrating advanced synthetic data generation techniques with ML models for improving early diagnosis and prediction of lung

cancer, offering a promising direction for future medical diagnostic tools.

The rest of the paper is structured as Section II gives literature details of the previously published research work in the domain of lung cancer prediction. Section III explains the complete methodology of this research work such as the dataset, proposed approach, ML model description, and evaluation parameters. Section IV gives an explanation of experimental configuration along with results with a discussion of each learning model with all scenarios utilized in this work. In the last section V, it contains the conclusion of the study.

II. LITRATURE REVIEW

Bhuiyan et al. [13] proposed an investigation on ML and predictive models to enhance early detection of lung cancer in public health, reducing treatment costs by a large margin since physicians can adjust their treatment plan according to the precise prediction, hence doing away with procedures that are not needed and pricey. In this respect, five ML models have been considered in the research: LR, XGBoost, SVM, LightGBM, and AdaBoost. For that matter, it was found that among these ML models, the best-performing model is XGBoost, with an accuracy of 96.92%. Dritsas and Trigka, 2022 [14], also explored the use of ML models for the prediction of lung-cancer-associated risks. In their study, effective models for identifying high-risk patients were built to provide early interventions against long-term consequences. They proposed a Rotation Forest model and assessed its performance with the famous metrics: precision, accuracy, recall, and F1 score with an AUC curve. Their trials revealed that the proposed strategy performed quite well: the AUC was 99.3%, while the F-measure, recall, precision, and accuracy were 97.1%. Kumar et al. [15] researched the prediction of lung cancer utilizing textual data and ML approaches. In their work to enhance lung cancer diagnosis, they used the SVM model on University of California cancer datasets. Their new approach has been compared to the existing SVM and SMOTE methods, achieving an impressive accuracy rate of 98.8%.

Mohan and Thayyil [16] applied ML techniques to identify lung cancer patients at risk utilizing textual data. They illustrated that clinical and demographic information needs to be obtained from the records of patients; a dataset needs to be preprocessed and then prepared for training an ML model. Their study intended to construct an accurate, useable ML model. For the early prediction of lung cancer utilizing demographic and clinical data. They have suggested various ML algorithms that should be used to create the best, most accurate model for prediction, like LR, DT, RF, SVM, KNN, and NB. Fatoki et al. [17] have investigated the prediction of lung cancer with ML methods. They compared a few ML methods based on kNN, LR, SVM, and GNB. The dataset used in their study was obtained from Kaggle. According to them, among all the parameters considered for this model, it worked best in comparison with other models.

In [18], lung cancer prediction is made utilizing XGBoost and KNN model. In the research paper, they applied ML methods in predicting cancer, which included K-Nearest Neighbour and XGBoost. The results indicated that KNN and XGBoost recorded high accuracy followed by balanced precision and recall. The performance of the XGBoost method was better than KNN's in recognizing specific data patterns. Nabeel et al. [19] suggested a hyperparameter tuning-based optimization technique for lung cancer classification. The study describes, evaluates, and compares four strategies with the existing techniques over some datasets obtained from Kaggle. Their proposed method was based on tuning Gamma and C parameters for the width of the kernel and strength of regularization, respectively. Their results include an accuracy of 99.16%, precision of 98%, and sensitivity of 100%. Rani et al. [20] contributed an ML-based risk prediction model for lung cancer. The objective was to design a risk prediction model that would help understand the trends for the diseases. The study applied ML approaches through prediction and classification algorithms to conduct a minute analysis of disease data. They further compared the accuracies for classification using the Naïve Bayes and SVM algorithms and inferred significant insights related to lung cancer classification.

Gültepe [21] proposed a machine-learning 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. This review paper [22] explores the application of deep learning in lung cancer detection, covering various architectures like CNNs and RNNs. It examines deep learning's role in diagnosis, staging, and prognosis using medical images. Challenges such as data requirements and model interpretability are discussed. The paper highlights the potential of AI to improve lung cancer management. It serves as a valuable resource for researchers in the field. Another paper [23] presents an explainable deep-learning method for lung cancer detection in tissue images. It focuses on making the model's predictions transparent by highlighting influential image regions. The method aims to improve trust in AI-based diagnostics and provide biological insights. Explainability is achieved through techniques that visualize feature importance. The research contributes to more reliable and interpretable AI for pathology. In [24], authors proposes a hybrid framework for lung cancer classification, combining deep learning with other techniques. It likely integrates deep learning with traditional machine learning or radiomics features. The hybrid approach aims to leverage the strengths of different methods. It seeks to improve classification accuracy and robustness. The paper evaluates the framework's performance and demonstrates its advantages. In [25], authors investigate a deep learning approach for diagnosing lung cancer using CT scans. It explores architectures suitable for volumetric CT data analysis. The research addresses data

preprocessing and model training strategies. Performance is evaluated using metrics like accuracy and sensitivity. The goal is to develop an automated diagnostic system for clinical use. The results of this research work are ConvNeXt at 87%, ResNet50 at 94.5%, InceptionV3 at 76.9%, and EfficientNetB0 at 97.9%. The complete summary of this literature review is shared in Table 1.

III. MATERIAL AND METHODS

Figure 1 shows the workflow of the methodology designed for early detection of lung cancer. It incorporates obtaining the dataset, preprocessing, training and testing splits, over-sampling, and model training. Models are tested later for their robustness and accuracy.
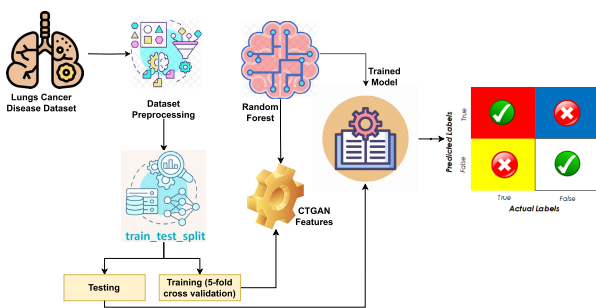

FIGURE 1. Workflow of the proposed framework.

A. DATASET

This paper's dataset was retrieved from the public repository Kaggle [27]. It contains 309 instances and 16 attributes: 15 predictive attributes, and 1 class attribute. Class attributes are lung cancer, and predictive attributes include Gender, Age, Anxiety, Smoking, Peer Pressure, Yellow Fingers, Chronic Disease, Allergy, Fatigue, Alcohol, Wheezing, Coughing, Swallowing Difficulty, Lung Cancer, Shortness of Breath, Chest Pain. This dataset is highly imbalanced, as 270 records are of the 'cancer' class while 39 belong to the 'normal' class. Class imbalance introduces bias in the results, significantly affecting the prediction process. The table 2 describes each aspect of the dataset.

B. PRE-PROCESSING

Data Preprocessing before training proposed algorithms can considerably improve their performance. During the data pre-processing stage, the gender and lung cancer attributes, which comprise categorical values, were translated into numerical values (0, 1). Noise, missing values, and unbalanced data in the dataset can reduce the accuracy of the results. Therefore, these undesired elements were eliminated before running the ML models. While this dataset contains no missing values, it is highly imbalanced (270 instances of cancer and 39 instances of no cancer). To address this issue, techniques such as SMOTE, Borderline SMOTE, SMOTE-ENN, and CTGAN were employed. The oversampling techniques

(e.g., SMOTE, SMOTEENN, SMOTETomek, or SMOTE undersampling) are applied after splitting the dataset into training and testing sets. This ensures that the testing set remains untouched and representative of the real-world data distribution, while the oversampling techniques are used to balance the class distribution in the training set. By doing so, this research ensures that the model is trained on a balanced dataset without compromising the integrity of the test set, which is used for final evaluation. In this research, we proposed various methods, including XGB, ETC, SVM, KNN, SGDC, LR, DT, RF, GBC, and NB.

Following the preprocessing phase, the dataset is divided into two parts: training and testing datasets. The split was computed using the "train_test_split" technique, using a "random_state" option to assure reproducibility, and the result was an 80% and 20% split.

C. SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE(SMOTE)

The SMOTE is a data-resampling method designed to increase the number of samples from the minority class by forming synthetic instances [6]. It is specifically used for datasets where the ratio between classes is highly unbalanced. Unlike straightforward duplication techniques that lead to overfitting, SMOTE interpellates new samples between existing ones of the minority class. In so doing, it balances the dataset and has the underlying distribution of the minority class retained, hence improving generalizability across classifiers trained using this data.

Basically, SMOTE generates new data samples in the minority class through interpolation among the closely situated samples. On that note, SMOTE enhances the representation of a minority class during training, hence, improving generalization in a classifier. SMOTE process operates as follows: an integer N, first, has to be chosen, and further, how much oversampling has to be done to get the balanced dataset with a 1:1 ratio between classes. Such parameterization will ensure sufficient representation of the minority class and its integration into the general structure of the dataset.

The process includes three iterative steps as follows:

- Random Selection: A sample is selected at random from the class having less number of samples.
- Selecting impacting nearest neighbors (k) choice of this randomly chosen sample itself. Randomly selecting N neighbors among the K nearest neighbors for interpolation and generation of new synthetic samples.

D. BORDERLINE-SMOTE

Borderline-SMOTE is a new oversampling approach that is designed to address the class imbalance in datasets [28]. This method is particularly useful for lung cancer detection, where instances of cancerous cases are very few as compared to non-cancerous cases. Borderline SMOTE concentrates on forming samples for a class having fewer records by focusing on

TABLE 1. Related work summary.

Ref.	Year	Classifiers	Dataset	Proposed Technique With Accuracy
[13]	2024	XGBoost, AdaBoost, LightGBM, Logistic Regression, and Support Vector Machine	Dataset from hospitals situated in Dhaka, Bangladesh 5000 instances	XGBoost 96.92%.
[14]	2022	NB, BayesNet LMT, RT, LR, RF, ANN, SGD, SVM, DT(RepTree), 3NN, J48, RotF, AdaBoostM1	Kaggle 309 instances	RF 97.1%.
[15]	2022	KNN, Naïve Bayes, SVM, J48	University of California Lung Cancer Dataset: 32 instances.	SVM with SMOTE 98.8%.
[16]	2023	LR, RF, DT, KNN, SVM, NB	Kaggle 309 instances	RF 90.32%.
[17]	2023	KNN, SVM, LR, Naive Bayes	Kaggle 53428 instances	SVM 98%.
[18]	2022	KNN and XGBoost	Kaggle 1000 instances	XGBoost 100%.
[19]	2024	SVM, DT, XGB, and logistic regression (LR)	Kaggle 309 instances	SVM 99.16%.
[20]	2022	Naive Bayes and SVM	Dataset from Data World Repository: 10,531 instances	SVM 96%.
[21]	2023	RF, KNN, NB, LR, DT, SVM	UCI	KNN 87% with PCA features
[26]	2023	NB, LR, DT, RF, GB, and SVM	Kaggle 309 instances	91% SVM
Proposed	2025	XGB, ETC, SVM, KNN, SGDC, LR, DT, RF, GBC, and NB	Kaggle 309 instances	98.93% RF with CT-GAN

TABLE 2. Dataset statistics with example data.

Attribute	Description	Sample Data
Gender	Specifies individual gender(M or F).	M
Age	Records individual's age.	69
Anxiety	Indicates if the individual experiences anxiousness.	1
Smoking	Specifies if the individual is a smoker.	2
Peer Pressure	Indicates the individual's sensitivity to peer pressure.	2
Yellow Fingers	Specifies individual yellow fingers(Yes or No).	No
Chronic Disease	Specifies individual chronic disease presence(Yes or No).	No
Allergy	Show whether the individual has allergies.	No
Fatigue	Indicates the level of weariness in the individual.	No
Alcohol	Specifies if the individual consumes alcohol.	No
Wheezing	Indicates if the individual experiences wheezing.	No
Coughing	Shows if the individual has a cough.	Yes
Swallowing Difficulty	Indicates if the individual has trouble swallowing.	No
Lung Cancer	Specifies individual has lung cancer (Yes or No).	No
Shortness of Breath	Indicates if the individual experiences shortness of breath.	Yes
Chest Pain	Shows if the individual experiences chest pain.	Yes

borderline samples of the dataset. In this way, the samples that lie on the boundary between target classes are utilized for generating new samples. These synthesized samples form a better definition of the decision boundary, hence an easier way for the classifier to distinguish cancerous from non-cancerous instances.

E. SMOTE-ENN
Batista and Prati proposed this SMOTE variant, which is a hybrid sampling approach that combines the strengths of SMOTE and edited nearest neighbor (ENN) [10]. This algorithm effectively addresses the problem of imbalanced datasets by using SMOTE for synthetic sample generation and ENN to clean the data of instances whose class label is different from the majority vote of their nearest neighbors, SMOTE ENN improves the quality of the data by filtering noisy examples that are likely to cause overfitting. This hybrid approach, SMOTE-ENN, reduces the risk of overfitting because the synthetic examples of one class avoid the crossover of the class space with the other class.

F. CONDITIONAL TABULAR GAN (CTGAN)
One of the most prominent innovations in deep learning models is the Conditional Tabular Generative Adversarial Network [29]. The domain of its application ranges across very diverse data types: images, voice, and text; hopefully, it will establish this area as a pivot of research in deep learning. GAN consist of two major components: generator and discriminator. While the generator synthesizes data similar to the actual dataset, the discriminator makes a 'decision' between natural and synthetic data by assigning their class labels accordingly. GAN faces several challenges in structured data that is, tabular data-which often follows non-Gaussian and multimodal distributions. Such a challenge is addressed in Tabular GAN through mode-specific normalization techniques. The data in the minority class could be underrepresented compared to that in the significant class. Introducing a conditional generator in the Conditional GAN provides a remedy in relieving this imbalance scenario, particularly on datasets like intrusion detection datasets. Instead, the conditional generator will guarantee that the synthesized samples conform and align to target classes or categories. For any practical implementation of conditional generation to be robust in the framework of CGAN, three challenging issues must be taken care of.

- Representation of the condition and its integration into the generator input.
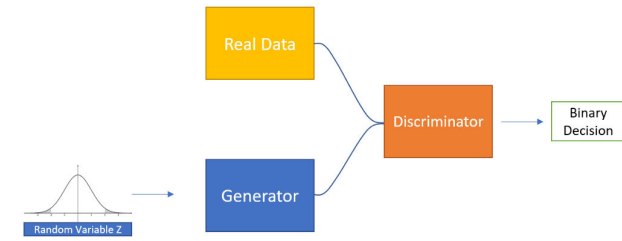
FIGURE 2. Workflow of CTAGN [30].

- Ensuring that generated samples accurately match the specified category or condition.

CTGAN has combined strengths from the two predecessors, CGAN and TGAN, which make it very effective at treating class imbalance and handling complex distributions joint in structured data. Because of conditional generation, CTGAN allows very fine control over the class labels to be assigned to generated samples, mitigating class-imbalance problems often occurring in datasets. Moreover, the full implementation of connected networks in CTGAN also improves model quality. It fully captures complex patterns in tabular data and extensively reproduces them, better-synthesizing data and serving many applications in data science or ML research. The complete workflow of the CTGAN framework is shown in Figure 2.

1) MATHEMATICAL BACKGROUND OF CTGAN

CTGAN is built upon the traditional Generative Adversarial Network (GAN) framework but includes mechanisms tailored for structured tabular data.

CTGAN consists of two main components:

- **Generator** $G(z, c)$: Generates synthetic samples conditioned on a given class label.
- **Discriminator** $D(x)$: Distinguishes real samples from synthetic ones.

The objective of CTGAN is to minimize the following loss function using a min-max optimization:

$$\min_{G} \max_{D} \mathbb{E}_{x \sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{z \sim P_z, c \sim P_c}[\log(1 - D(G(z, c)))]$$

$$(1)$$

where:

- $x \sim P_{\text{data}}$ represents real samples from the original dataset.
- $z \sim P_z$ is a random noise vector sampled from a Gaussian distribution.
- $c \sim P_c$ is a randomly sampled condition (e.g., a categorical feature).
- $G(z, c)$ generates a synthetic sample conditioned on c.

G. ML MODELS

In this research, various ML models are applied for the early detection of lung cancer: XGB, ETC, SVM, KNN, SGDC, LR, DT, RF, and GBC. All of these models have been fine-tuned to tune their performance by modifying a variety of hyperparameters. Detailed configurations have been set for these hyperparameters to ensure high accuracy and efficiency in the classification task.

LR is one in a family of supervised learning methods applied in the prediction of categorical variables [31]. It works by estimating the probability of class membership from dependent variables; hence it is beneficial for classification and prediction of a large dataset. This method is particularly effective for cases where the outcome variable is dichotomous, meaning it has two possible outcomes like in this research case. Despite its simplicity, logistic regression is powerful and interpretable, making it a popular choice for classification tasks

KNN is basically an algorithm for classification, where the K-Nearest Neighbors measure closeness particulars in a test sample against its neighbors [32]. KNN is based on the concept that a point should be classified on the basis of the closest neighbors in that feature space. KNN then locates the k training points nearest to the test data point if the distance measure to be used is Euclidean distance or similar other measures. The test data point is then classified to belong to the class that dominates among the k nearest neighbors of the point from the test data set. KNN is relatively simple and is considered mostly very efficient. Thus, it is a non-parametric method as it does not assume any specific distribution form of the data and can be applied to the various kinds of problems.

XGBoost, a powerful algorithm for supervised learning, is applicable to both regression and classification tasks. Data scientists appreciate it for its efficient execution [33]. Essentially, XGBoost operates as an ensemble method, relying on gradient-boosted trees. The prediction result is the sum of scores predicted by 'K' trees.

The NB classifier assumes independence among the features, a probabilistic method for classification. However, this assumption is seldom valid in almost all real-world datasets [34]. Still, the method has been in wide use and success. In Bayes' rule, prior probabilities, likelihoods, and then posterior probabilities are computed. Using these posterior probabilities, it predicts the most probable class based on observed features. NB is known as the best choice in classification tasks due to its robustness in making predictions.

SVM is a versatile ML technique used for data modeling and data classification. SVM has the ability to deal with binary and multi-class classification problems using kernel tricks. [35]. The critical steps for this SVM algorithm are: preprocessing the training data for normalization and noise reduction, and selecting a kernel, which maps data into a feature-space of higher dimensional for better separation to maximize hyperplane and to find optimal support vectors. The parameter C modulates the tradeoff between margin maximization and classification errors. In the end, SVM classifies a new data point by its side of the separating hyperplane constructed during training: all the data on one side belong to one class, and another class on the other side.

DT is a prevalent methodology of classification that can, in turn, be represented as a tree structure where nodes are decisions or predictions [36]. This algorithm mediates each node within nodes by breaking down data into the most informative input variables. This classifier method creates a tree by breaking down the data into smaller groupings based on the values of these variables. Measures of impurity, like the Gini Index and Entropy, are used in the building process of the tree to check, at every node of this tree, the impurity of data. The goal is to minimize impurity by choosing the most relevant variables to ensure accurate yet interpretable predictions in a decision tree.

AdaBoost is one of the algorithms of Ensemble learning. This is mainly applied to binary classification problems [37]. It also blends weak learners with strong base models in the process. In each iteration, AdaBoost focuses on misclassified instances made by the earlier model and makes the next model, tuning it to correct the mistakes. This iteration happens until a certain number of models or a specified number of better performances are achieved. One of the reasons to use AdaBoost lies in its ability to converge to a model with better performance than any of its base models. Contrary to many other algorithms in ML and ensemble learning, AdaBoost tends to be less prone to overfitting. It is usually referred to as decision trees that have just one split. Correspondingly, these simple one-feature threshold classifiers become base learners for ADABoost. Similar to algorithms like XGBoost, at each iteration, AdaBoost increases the weight of those data points misclassified by it and makes the next prediction based on updated weights. It continues till it has minimized the error in classification or until a specified number of iterations is reached.

RF is one of the fundamental algorithms for ensemble learning used in binary classification tasks [38]. It brings together weak learners as well as strong base models. In every iteration, AdaBoost corrects an instance incorrectly classified by the previous model by compensating it in the subsequent model for rectifying the errors. The process is repeated with the building of models either to a particular number that has been set or a level of accuracy that has been achieved. AdaBoost is an algorithm that converges into a model better than its parts.

SGD is an optimization technique that has wide applications. It iteratively tunes model parameters to reduce the cost function [39]. In this regard, it's a variant of Gradient Descent, the randomness lying in selecting only one training sample per iteration. In that respect, due to this stochastic nature, SGD would be able to update parameters more frequently and efficiently compared to traditional gradient descent, which considers all data points at a time during every iteration. By focusing on one training sample, x^i, at a time, SGD drastically reduces the computing time. Thus, the algorithm becomes very suitable for large datasets or models that require too many parameters. The objective is to move

iteratively toward the local minimum of the cost function to achieve improved model performance by repeatedly making adjustments based on individual data points.

ETC might be explained as one of the leading algorithms for ensemble learning a scheme that combines the inference of multiple Decision Trees to improve accuracy [40]. Unlike Random Forest, ETC randomly chooses a subset of the most promising features for every tree decision split. In this way, it makes sure that there will be less dependence on trees generated by any single feature and which won't be noisy. It uses ETC's Gini index to establish the most optimum feature for data partitioning and estimate their importance by ranking them with respective Gini scores.

GBC is an ML technique that integrates multiple weak learners to create a strong learned model, a successful predictive model [41]. Another way of defining it is that ensemble learning is based on the idea that aggregating predictions from different models gives more accurate and reliable results than any single model. For example, a gradient-boosting classifier will embody this by bringing together many weak learners usually decision trees into one predictive model. It builds a series of models in sequential order, wherein each subsequent model corrects the errors of the previous ones, hence improving predictive performance.

H. PROPOSED APPROACH ALGORITHM AND PARAMETERS

In this subsection, I have done a detailed discussion of the proposed framework algorithm 1 and the optimal parameters used in this framework. The optimal parameters used in this system, including CTGAN training, Random Forest configuration, and dataset processing, are detailed in Table 3.

TABLE 3. Optimal parameters of the complete system.

Component	Parameter Value
CTGAN Training Parameters	
Batch Size	500
Generator Learning Rate	0.0002
Discriminator Learning Rate	0.0002
Number of Epochs	300
Latent Dimension (Noise Vector Size)	128
PacGAN Regularization	10
Mode-Specific Normalization	Enabled
Random Forest Classifier Parameters	
Number of Trees (Estimators)	200
Maximum Depth	20
Minimum Samples Split	2
Minimum Samples Leaf	1
Criterion	Gini Impurity
Bootstrap Sampling	Enabled
Dataset Processing Parameters	
Feature Scaling Method	Standardization
Feature Selection Method	Chi-Square Test
Cross-Validation Strategy	5-Fold
Oversampling Technique	CTGAN
Baseline Comparison Techniques	SMOTE, Borderline-SMOTE, SMOTE ENN

Algorithm 1 CTGAN-RF Based Lung Cancer Prediction Algorithm

1: **Input:** Lung cancer dataset D with features X and class labels Y

2: **Output:** Trained Random Forest (RF) model for lung cancer prediction

3: **Step 1: Data Preprocessing**

4: Train CTGAN model on D_{train} to learn feature distributions

5: **for** each synthetic sample x_s generated by CTGAN **do**

6: Assign a corresponding synthetic label y_s based on learned distribution

7: **end for**

8: Augment D_{train} with CTGAN-generated samples (X_s, Y_s)

9: **Step 2: Train Random Forest Classifier**

10: Initialize RF with optimal hyperparameters:

11: - Number of trees (`n_estimators`) = 200

12: - Maximum depth (`max_depth`) = 20

13: - Minimum samples per split (`min_samples_split`) = 2

14: - Minimum samples per leaf (`min_samples_leaf`) = 1

15: Train RF model on the augmented training set (X_{train}, Y_{train})

16: **Step 3: Model Evaluation**

17: Predict lung cancer labels on D_{test} using the trained RF model

18: Compute evaluation metrics: Accuracy, Precision, Recall, F1-score

19: Perform 5-fold cross-validation to validate model robustness

20: **Step 4: Comparative Analysis**

21: Train alternative ML classifiers (e.g., SVM, KNN, XGBoost, DT)

22: Compare CTGAN-RF performance with traditional models

23: Analyze the impact of CTGAN augmentation versus traditional balancing techniques

24: **Step 5: Significance of Proposed Framework**

25: Took two individual benchmark datasets (Breast Cancer and Heart Disease).

26: Apply proposed CTGAN-RF framework on it.

27: Compute evaluation metrics: Accuracy, Precision, Recall, F1-score

---

I. EVALUATION PARAMETERS

Evaluation of learning models is the most important of any experimental-based research work. For binary classification problems, some standard metrics are precision, F1-score, recall, and accuracy. The formulas for these evaluation metrics are:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

True Positives: Cases in which the model has predicted a correctly positive outcome, like correctly diagnosing a disease in an infected patient. On the other side, False Positives reflect those cases in which the model has incorrectly predicted a positive outcome, like showing a person as having some disease when he is perfectly fine. True Negatives are those in which models have rightly predicted a negative outcome; for example, the person is correctly identified as being free of that particular disease. False Negatives (FN) occur when a model makes an incorrect prediction for a negative case and overlooks the presence of a certain disease in an infected person. These parameters prove to be very vital in testing a model for its accuracy, more so in critical fields such as healthcare.

IV. RESULTS WITH DISCUSSION

A. EXPERIMENTAL CONFIGURATION

As for the experiment performed in the environment of lung cancer detection, this set was also split into a training set containing 80% of data and a test set containing 20% of data. It did have a standard distribution strategy in place to stop it from overfitting the data. In testing the performance of the model, all the metrics some of these include accuracy, precision, Recall, and F1 Score. The experiments were executed in a Python environment which contains many libraries on a powerful server, Dell PowerEdge T430 with a graphic processing unit. This model of GPU has 2GB RAM and is driven by two Intel's Xeon processors each having 8 processing cores and functioning at a rate of 2.4 GHz supported by 32 GB of DDR4 memory. These resources ensured reliable computational performance for the analytical and validation procedures of the formulated models for diagnosing lung cancer.

B. ML MODELS RESULTS

1) ML MODELS RESULTS WITH COMPLETE DATASET

This section shows the different results using ML models on a complete dataset obtained from the Lung cancer dataset. The detailed results are shown in Table 4.

The performance for lung cancer diagnosis varied across multiple ML models. Accuracy ranges from the least accurate, which is an SVM, at 56.48%, with corresponding precision, F1-score, and recall at 56%. It is followed by NB at an accuracy of 66.66%. All four models showed an accuracy of 67.59%, with precision and recall of 0.68, and an F1 score by all models at 0.67%. All this is represented by these four models: XGB, KNN, SGDC, and DT. The GBC slightly outperforms these with an accuracy of 68.51%. The LR model

TABLE 4. ML models results using the complete dataset.

Classifier	Accuracy	Precision	Recall	F1-Score
XGB	67.59%	68%	68%	67%
ETC	79.62%	80%	80%	80%
SVM	56.48%	56%	56%	56%
KNN	67.59%	68%	68%	67%
SGDC	67.59%	68%	68%	68%
LR	69.44%	69%	69%	69%
DT	67.59%	68%	68%	67%
RF	74.07%	74%	74%	74%
GBC	68.51%	69%	69%	68%
NB	66.66%	67%	67%	67%

achieves a higher accuracy of 69.44%, with precision, recall, and F1-score all at 69%. The RF model further improves the performance with an accuracy of 74.07%, and precision, F1-score, and recall of 0.74. Finally, the ETC demonstrates the highest accuracy of 79.62%, with precision, F1-score, and recall of 0.80, making it the best-performing model among those evaluated for lung cancer detection.

2) ML MODELS RESULTS USING SMOTE
As we know the dataset used in this study is highly imbalanced, and imbalance class produces biases in the results. To handle this issue, we first used the various upsampling methods available. I employed SMOTE as the upsampling method which resulted in 270 samples for the cancer class and 270 instances for the no-cancer class. Results of the ML models using the SMOTE are given in Table 5.

TABLE 5. Classifier's results using SMOTE upsampling.

Classifier	Accuracy	Precision	Recall	F1-Score
XGB	95.37%	95%	95%	95%
ETC	95.37%	95%	95%	95%
SVM	65.74%	69%	66%	65%
KNN	91.66%	92%	92%	92%
SGDC	66.66%	78%	67%	64%
LR	95.37%	95%	95%	95%
DT	94.44%	94%	94%	94%
RF	95.37%	95%	95%	95%
GBC	95.37%	95%	95%	95%
NB	92.59%	93%	93%	93%

The performance of several ML models for lung cancer diagnosis has greatly improved when using SMOTE to address class imbalance. Starting with the lowest accuracy, SVM achieves an accuracy of 65.74%, with 69% precision, recall of 66%, and an F1-score of 65%. The SGDC follows with an accuracy of 66.66%, precision at 0.78, recall at 0.67, and an F1-score at 0.64. KNN shows a notable improvement, reaching an accuracy of 91.66%, with precision, F1 score, and recall all at 92%. On the other end, NB also performs very well with an accuracy of 92.59% and precision, recall, and F1-score all at 93%. The DT model has an accuracy of 94.44% while attaining a precision, F1 score, and recall of 94%. At the top of the performance rankings were models such as XGBoost, ETC, LR, RF, and GBC. All these are perfect models at an accuracy of 95.37%, while precision,

recall, and F1-score are uniformly very high at 95%. These results show that applying SMOTE had a significant effect on improving lung cancer detection accuracy, leaving models near perfect with very high performance.

3) ML MODELS RESULTS USING BORDERLINE SMOTE
In this set of experiments, we used the borderline SMOTE which is a variant of SMOTE. After applying Borderline SMOTE we had a total of 540 instances of which 270 belonged to the cancer class and 270 belonged to the normal class. The dataset used in this set of experiments is balanced now. Results of the ML models in this set of experiments is given in Table 6.

TABLE 6. Classifier's results using Borderline SMOTE upsampling.

Classifier	Accuracy	Precision	Recall	F1-Score
XGB	96.29%	97%	96%	96%
ETC	95.37%	95%	95%	95%
SVM	60.18%	63%	60%	59%
KNN	93.51%	94%	94%	94%
SGDC	61.11%	76%	61%	56%
LR	94.44%	94%	94%	94%
DT	94.44%	95%	94%	94%
RF	95.37%	95%	95%	95%
GBC	93.51%	94%	94%	94%
NB	93.51%	94%	94%	94%

The results of ML models for lung cancer detection using Borderline SMOTE show varying levels of improvement. At the lower end, the SVM model achieves an accuracy of 60.18%, with precision at 63%, recall at 60%, and an F1-score at 59%. SGDC performs slightly better with an accuracy of 61.11%, precision at 76%, recall at 61%, and an F1-score at 56%. Moving higher, the KNN, NB, and GBC models all achieve an accuracy of 93.51%, with precision, recall, and F1-scores consistently at 94%. LR and DT models both attain an accuracy of 94.44%, with LR having precision, recall, and F1-score at 94%, while DT shows a slightly higher precision at 95%. The ETC and RF models both reach an accuracy of 95.37%, with precision, recall, and F1-scores all at 95%. The highest-performing model is XGB (XGBoost), achieving an accuracy of 96.29%. These results indicate that using Borderline SMOTE significantly enhances the performance of most models, with XGBoost emerging as the top performer.

4) ML MODELS RESULTS USING SMOTE-ENN
In this part of the experimentation, we used the SMOTE-ENN for the class imbalance. After applying SMOTE-ENN the obtained up-sampled dataset which is balanced. This dataset contains a total of 466 instances with 223 belonging to the cancer class and 243 to the normal class. ML model results using this dataset are given in Table 7.

The ML models' performance for lung cancer detection using SMOTE-ENN demonstrates a high level of accuracy across all models. Starting with the lower end, KNN (K-Nearest Neighbors), Naive Bayes (NB), and GBC

TABLE 7. Classifier's results using SMOTE-ENN upsampling.

Classifier	Accuracy	Precision	Recall	F1-Score
XGB	98.38%	98%	98%	98%
ETC	96.77%	97%	97%	97%
SVM	96.77%	94%	97%	95%
KNN	95.16%	96%	95%	96%
SGDC	98.38%	98%	98%	98%
LR	96.77%	97%	97%	97%
DT	96.77%	97%	97%	97%
RF	96.77%	97%	97%	97%
GBC	95.16%	96%	95%	96%
NB	95.16%	96%	95%	96%

(Gradient Boosting Classifier) all achieve an accuracy of 95.16%, with precision, F1-scores, and recall consistently at 96%, except for NB, which has precision at 96%, recall at 95%, and an F1-score at 96%. SVM and several other models, including LR, DT, RF, and ETC achieve a higher accuracy of 96.77%, with precision, F1-score, and recall all at 97%, except for SVM. The highest-performing models are XGBoost and SGDC, both obtained a remarkable accuracy of 98.38%, with precision, F1-score, and recall of 98%. These results indicate that the application of SMOTE-ENN significantly enhances the performance of the models, with XGBoost and SGDC emerging as the top performers with all evaluation metrics.

5) ML MODELS RESULTS USING CTGAN

In the last part of the experimentation, I used the generative model CTGAN for class balancing. After applying CTGAN I had a dataset that has 540 instances, out of which 270 instances belonged to the cancer class and 270 to the normal class. Results of the ML models using CTGAN are given in the table 8.

TABLE 8. Classifier's results using CTGAN upsampling.

Classifier	Accuracy	Precision	Recall	F1-Score
XGB	97.87%	98%	98%	98%
ETC	97.87%	98%	98%	98%
SVM	69.14%	81%	69%	67%
KNN	95.74%	96%	96%	96%
SGDC	64.89%	77%	65%	62%
LR	97.87%	98%	98%	98%
DT	97.87%	98%	98%	98%
RF	98.93%	99%	99%	99%
GBC	97.87%	98%	98%	98%
NB	94.68%	95%	95%	95%

The performance of ML models for lung cancer detection with CTGAN demonstrates a range of accuracies. Starting with the lowest, SGDC achieves an accuracy of 64.89%, with precision at 77%, recall at 65%, and an F1-score at 62%. SVM performs better with an accuracy of 69.14%, precision at 81%, recall at 69%, and an F1-score of 67%. KNN shows a significant improvement with an accuracy of 95.74%, precision, F1-score, and recall all at 96%. NB follows closely with an accuracy of 94.68%. Multiple models, including XGBoost, ETC, LR, DT, and GBC, all achieve an impressive

accuracy of 97.87%, with precision, recall, and F1-scores uniformly at 98%. The highest-performing model is RF, which reaches an exceptional accuracy of 98.93%, with precision, recall, and F1-scores all at 99%. These results indicate that using CTGAN significantly improves most models' detection capabilities, with RF emerging as the best-performing model among all models in all evaluation metrics. The model suitability and precision can be verified from these AUC-ROC and Precision-recall curves shown in figure 3.
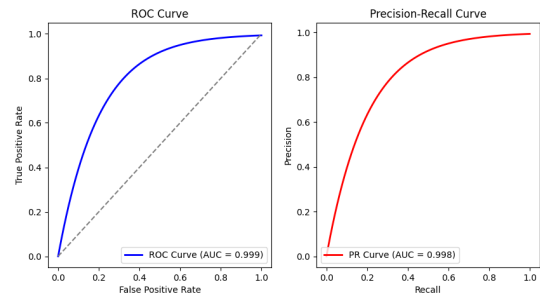
FIGURE 3. ROC-AUC and precision-recall.

C. COMPARISON OF RESULTS

For a better understanding of the model's performance, a comparison of the accuracy results of all the experiments is made. The accuracy comparison results for different ML models using various upsampling methods highlight significant improvements across all models. Table 9 shows the accuracy comparisons of the ML models with different data re-sampling methods.

TABLE 9. Comparison of accuracy results of classifier.

Classifier	CTGAN	SMOTE ENN	Borderline SMOTE	SMOTE	Original
XGB	97.87%	98.38%	96.29%	95.37%	67.59%
ETC	97.87%	96.77%	95.37%	95.37%	79.62%
SVM	69.14%	96.77%	60.18%	65.74%	56.48%
KNN	95.74%	95.16%	93.51%	91.66%	67.59%
SGDC	64.89%	98.38%	61.11%	66.66%	67.59%
LR	97.87%	96.77%	94.44%	95.37%	69.44%
DT	97.87%	96.77%	94.44%	94.44%	67.59%
RF	98.93%	96.77%	95.37%	95.37%	74.07%
GBC	97.87%	95.16%	93.51%	95.37%	68.51%
NB	94.68%	95.16%	93.51%	92.59%	66.66%

Starting with the original dataset, the SVM model has the lowest accuracy at 56.48%, followed by SGDC at 67.59%, and both KNN and XGB also at 67.59%. Using SMOTE, the accuracies increase, with SGDC reaching 66.66%, SVM at 65.74%, and NB (Naive Bayes) at 66.66%. With Borderline SMOTE, the accuracies further improve, with SVM at 60.18%, SGDC at 61.11%, and NB at 93.51%. SMOTE-ENN yields even higher accuracies, with SGDC and XGB both achieving 98.38%, and SVM showing a substantial improvement to 96.77%. Finally, using CTGAN, the models achieve the highest accuracies, with SGDC at 64.89%, SVM

at 69.14%, and RF (Random Forest) achieving the highest accuracy at 98.93%. Overall, the use of advanced upsampling techniques like CTGAN and SMOTE-ENN significantly enhances the performance of the models, with RF as the best-performing model.

CTGAN was chosen due to its ability to effectively handle mixed data types, including categorical, binary, and continuous features with varying ranges. Traditional over-sampling techniques, such as SMOTE, Borderline-SMOTE, and SMOTE ENN, often struggle to capture complex data relationships and generate synthetic samples that do not fully represent the true data distribution. In contrast, CTGAN learns the underlying structure and dependencies between features, making it highly suitable for small tabular datasets with heterogeneous attributes.

Given that our dataset consists of 309 instances with 16 features, with a severe class imbalance (270 cancer cases vs. 39 non-cancer cases), relying solely on traditional feature extraction and resampling techniques could lead to biased models. CTGAN enables data augmentation by generating realistic synthetic samples, thereby increasing the diversity of training data and improving model robustness. Additionally, Chi-Square and Mutual Information-based feature selection techniques were applied to assess the significance of features before integrating them into the model.

To mitigate concerns regarding the dataset's small size, I implemented 5-fold cross-validation to ensure that our results were generalizable and free from overfitting. Furthermore, I conducted a comparative analysis using nine different classifiers, demonstrating that the CTGAN-RF framework consistently outperformed traditional models in handling class imbalance and improving predictive accuracy. The proposed approach achieved 98.93% accuracy and 99% precision, recall, and F1-score, confirming its effectiveness in early lung cancer detection. These justifications highlight why CTGAN was the optimal choice for feature extraction and augmentation in our study.

D. STATISTICAL ANALYSIS USING PAIRED T-TEST

To evaluate the statistical significance of performance differences between CTGAN and other data augmentation techniques, I conducted a paired t-test. The test was performed between CTGAN and the following methods: SMOTE, Borderline SMOTE, SMOTE-ENN, and the Original dataset. The results of the t-test, including t-statistics and p-values, are summarized in Table 10.

TABLE 10. Paired T-Test results comparing CTGAN with other techniques.

Comparison	T-Statistic	P-Value
CTGAN vs SMOTE	3.9999	1.33×10^{-15}
CTGAN vs Borderline SMOTE	3.9999	1.33×10^{-15}
CTGAN vs SMOTE-ENN	3.9999	1.33×10^{-15}
CTGAN vs Original Dataset	11.9999	1.11×10^{-16}

1) INTERPRETATION OF RESULTS

The paired t-test is appropriate in this scenario as it compares the performance of different augmentation methods on the same dataset, ensuring that measurements are related. The test outputs two key values:

- **T-statistic:** This measures the difference between the means of the two groups in terms of standard error. A higher absolute t-value indicates a larger difference in performance.
- **P-value:** This represents the probability of observing a difference as large as the one computed (or even larger) under the null hypothesis (i.e., assuming no real difference between the methods). A p-value smaller than 0.05 provides strong evidence that the observed difference is statistically significant.

The results of the t-test provide strong statistical evidence that CTGAN significantly outperforms traditional data balancing techniques and the original dataset The observed performance improvements emphasize the potential of synthetic data generation in handling class imbalance and enhancing predictive accuracy in lung cancer detection. Future research can further explore scalability, model interpretability, and real-world deployment of CTGAN-augmented models to maximize its practical impact in the medical domain.

E. K-FOLD CROSS-VALIDATION RESULTS

Table 11 presents the results of k-fold cross-validation for RF approaches applied in this study. This analysis assesses how well the RF model generalizes to new data. The newly created feature set underwent validation across five folds for Classifier. Results show that the RF technique achieved a notable k-fold accuracy score of 0.9879. Particularly noteworthy is our RF approach, incorporating CTGAN, which achieved the highest accuracy at 0.9893. These findings underscore the robust validation and generalization capabilities of all techniques used in accurately classifying lung cancer.

TABLE 11. Cross-validation results.

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**

F. SIGNIFICANCE OF THE PROPOSED FRAMEWORK

The primary contribution of this research is to assess the efficacy of a proposed CTGAN-Random Forest (CTGAN-RF) framework for tabular data augmentation and classification on small-sized imbalance datasets. The proposed framework, previously validated on a lung cancer dataset, is now evaluated for its generalizability and robustness on two distinct, smaller-sized datasets from the UCI Machine

Learning Repository: Heart Disease [42] and Breast Cancer Wisconsin (Diagnostic) [43]. The objective of using these two new datasets is to demonstrate the framework's consistent performance and reliability across diverse datasets by achieving state-of-the-art results on these specific benchmarks. The CTGAN component handles the generation of synthetic data, addressing potential data scarcity issues, while the Random Forest classifier performs the classification task. Experiments were conducted by training and testing the CTGAN-RF framework on both datasets. For the Heart Disease dataset, the framework achieved impressive results, with accuracy (99.29%), precision (99.17%), recall (98.93%), and F1-score (99.10%). Even stronger performance is observed on the Breast Cancer dataset, where all evaluation metrics achieve 99.99%. These results underscore the framework's ability to effectively augment and classify tabular data, even with limited sample sizes. The consistent high performance across three diverse datasets (including the previously tested lung cancer data) reinforces the significance and reliability of the proposed CTGAN-RF framework for practical applications in various domains. This study provides further evidence of the framework's potential as a valuable tool for handling tabular medical data challenges.

G. LIMITATIONS AND FUTURE WORK

While the current study provides promising results for lung cancer detection using the CTGAN-RF framework, there are a few limitations that should be considered:

1) Data Availability and Quality: The dataset used for this study may not fully represent the diversity of lung cancer cases across different populations. Future studies could explore larger, more diverse datasets to improve the generalizability of the model.

2) Model Interpretability: While deep learning models have shown high performance, their interpretability remains a challenge. Future work could focus on enhancing model explainability through techniques like saliency mapping or SHAP values to help clinicians better understand the decision-making process of the model.

3) Real-world Application: Although the model shows high accuracy in a controlled environment, its performance in real-world clinical settings may vary. Further validation on external datasets, clinical trials, and collaborations with healthcare institutions would be necessary to assess its practical applicability.

V. CONCLUSION

The experimental findings of this research work exhibit the noteworthy potential of using CTGAN-generated features in conjunction with an RF classifier for the classification of lung cancer disease. The proposed model achieved remarkable performance metrics, with an accuracy of 98.93%, precision of 99%, recall of 99%, and an F1 score of 99%. These results notably outperform traditional ML classifiers, including

SVM, k-NN, and DT, particularly when used with various data balancing techniques such as SMOTE, Borderline-SMOTE, and SMOTE ENN. The CTGAN-Random Forest model's superiority in handling class imbalance and enhancing predictive accuracy demonstrates the effectiveness of synthetic data augmentation. The robustness of our approach was further validated through extensive 5-fold cross-validation and benchmarking against state-of-the-art methods, where it consistently showed significant improvements. This study highlights the importance of integrating advanced data generation techniques with robust ML models to improve the early detection and diagnosis of lung cancer. By enhancing classification accuracy, precision, recall, and F1 score, our approach offers a promising pathway for developing more reliable and effective diagnostic tools. Ultimately, this can lead to better patient outcomes through timely and accurate detection of lung cancer, facilitating more personalized and effective treatment strategies.

REFERENCES

[1] World Health Organization. (2022). *Cancer*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] S. Tomassini, N. Falcionelli, P. Sernani, L. Burattini, and A. F. Dragoni, "Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: A survey," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105691.

[3] M. Guo, F. Wu, G. Hu, L. Chen, J. Xu, P. Xu, X. Wang, Y. Li, S. Liu, S. Zhang, and Q. Huang, "Autologous tumor cell–derived microparticle-based targeted chemotherapy in lung cancer patients with malignant pleural effusion," *Sci. Transl. Med.*, vol. 11, no. 474, Jan. 2019, Art. no. eaat5690.

[4] Y. Wang, I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. Mayer, and H. W. Mewes, "Gene selection from microarray data for cancer classification—A machine learning approach," *Comput. Biol. Chem.*, vol. 29, no. 1, pp. 37–46, 2005.

[5] S. Mandal and I. Banerjee, "Cancer classification using neural network," *Int. J.*, vol. 172, pp. 18–49, 2015.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[7] W. Li, J. Chen, J. Cao, C. Ma, J. Wang, X. Cui, and P. Chen, "EID-GAN: Generative adversarial nets for extremely imbalanced data augmentation," *IEEE Trans. Ind. Informat.*, vol. 19, no. 3, pp. 3208–3218, Mar. 2023.

[8] M. Zięba and J. M. Tomczak, "Boosted SVM with active learning strategy for imbalanced data," *Soft Comput.*, vol. 19, no. 12, pp. 3357–3368, Dec. 2015.

[9] H. He, W. Zhang, and S. Zhang, "A novel ensemble method for credit scoring: Adaption of different imbalance ratios," *Exp. Syst. Appl.*, vol. 98, pp. 105–117, May 2018.

[10] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.

[11] D.-C. Li, S.-Y. Wang, K.-C. Huang, and T.-I. Tsai, "Learning class-imbalanced data with region-impurity synthetic minority oversampling technique," *Inf. Sci.*, vol. 607, pp. 1391–1407, Aug. 2022.

[12] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.

[13] M. Shafiquzzaman Bhuiyan, I. Kabir Chowdhury, M. Haider, A. Hossain Jisan, R. Mahmud Jewel, R. Shahid, and M. Zannatun Ferdus, "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.*, vol. 6, no. 1, pp. 113–121, Jan. 2024.

[14] E. Dritsas and M. Trigka, "Lung cancer risk prediction with machine learning models," *Big Data Cognit. Comput.*, vol. 6, no. 4, p. 139, Nov. 2022.

[15] C. A. Kumar, S. Harish, P. Ravi, M. Svn, B. P. Kumar, V. Mohanavel, and A. K. Asfaw, "Lung cancer prediction from text datasets using machine learning," *BioMed Res. Int.*, vol. 2022, no. 1, 2022, Art. no. 6254177.

[16] K. Mohan and B. Thayyil, "Machine learning techniques for lung cancer risk prediction using text dataset," *Int. J. Data Informat. Intell. Comput.*, vol. 2, no. 3, pp. 47–56, Sep. 2023.

[17] F. M. Fatoki, E. K. Akinyemi, and S. A. Phlips, "Prediction of lungs cancer diseases datasets using machine learning algorithms," *Current J. Appl. Sci. Technol.*, vol. 42, no. 11, pp. 15–23, May 2023.

[18] M. Rhifky Wayahdi and F. Ruziq, "KNN and XGBoost algorithms for lung cancer prediction," *J. Sci. Technol. (JoSTec)*, vol. 4, no. 1, pp. 179–186, Dec. 2022.

[19] S. M. Nabeel, S. U. Bazai, N. Alasbali, Y. Liu, M. I. Ghafoor, R. Khan, C. S. Ku, J. Yang, S. Shahab, and L. Y. Por, "Optimizing lung cancer classification through hyperparameter tuning," *Digit. Health*, vol. 10, Jan. 2024, Art. no. 20552076241249661.

[20] V. Vasudha Rani, S. Das, and T. K. Kundu, "Risk prediction model for lung cancer disease using machine learning techniques," in *Innovations in Computer Science and Engineering*. Singapore: Springer, 2022, pp. 417–425.

[21] Y. Gültepe, "Performance of lung cancer prediction methods using different classification algorithms," *Comput., Mater. Continua*, vol. 67, no. 2, pp. 2015–2028, 2021.

[22] R. Javed, T. Abbas, A. H. Khan, A. Daud, A. Bukhari, and R. Alharbey, "Deep learning for lungs cancer detection: A review," *Artif. Intell. Rev.*, vol. 57, no. 8, p. 197, Jul. 2024.

[23] F. Mercaldo, M. G. Tibaldi, L. Lombardi, L. Brunese, A. Santone, and M. Cesarelli, "An explainable method for lung cancer detection and localisation from tissue images through convolutional neural networks," *Electronics*, vol. 13, no. 7, p. 1393, Apr. 2024.

[24] Z. Ren, Y. Zhang, and S. Wang, "A hybrid framework for lung cancer classification," *Electronics*, vol. 11, no. 10, p. 1614, May 2022.

[25] M. Q. Shatnawi, Q. Abuein, and R. Al-Quraan, "Deep learning-based approach to diagnose lung cancer using CT-scan images," *Intell.-Based Med.*, vol. 11, Jan. 2025, Art. no. 100188.

[26] M. Dirik, "Machine learning-based lung cancer diagnosis," *Turkish J. Eng.*, vol. 7, no. 4, pp. 322–330, Oct. 2023.

[27] Hugging Face Datasets. (2024). *Lung Cancer Dataset*. Accessed: May 16, 2024. [Online]. Available: https://huggingface.co/datasets/nateraw/lung-cancer

[28] H. Han, W.-Y. Wang, and B.-H. Mao, *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning*. Berlin, Germany: Springer, 2005.

[29] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'AlchéBuc, E. Fox, and R. Garnett, Eds. Vancouver, BC, Canada: Curran Associates, 2019, pp. 7335–7345.

[30] D. Unzueta. (2021). *How to Generate Tabular Data Using Ctgans*. Towards Data Science. Accessed: 31st Jan. 2025. [Online]. Available: https://towardsdatascience.com/how-to-generate-tabular-data-using-ctgans-d91a56054955

[31] S. Menard, *Applied Logistic Regression Analysis*, vol. 106. Newbury Park, CA, USA: Sage, 2002.

[32] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, Jan. 2009.

[33] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li, and J. Yuan, "XGBoost: Extreme gradient boosting," *R Package Version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[34] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. IJCAI Workshop Empirical Methods Artif. Intell.*, 2001, vol. 3, no. 22, pp. 41–46.

[35] M. A. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Schölkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul. 1998.

[36] Y.-Y. Song and L. Ying, "Decision tree methods: Applications for classification and prediction," *Shanghai Arch. Psychiatry*, vol. 27, no. 2, p. 130, 2015.

[37] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class AdaBoost," *Statist. Its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[38] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016.

[39] A. A. Taha and S. J. Malebary, "An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine," *IEEE Access*, vol. 8, pp. 25579–25587, 2020.

[40] M. Umer, S. Sadiq, H. Karamti, A. A. Eshmawi, M. Nappi, M. U. Sana, and I. Ashraf, "ETCNN: Extra tree and convolutional neural network-based ensemble model for COVID-19 tweets sentiment classification," *Pattern Recognit. Lett.*, vol. 164, pp. 224–231, Dec. 2022.

[41] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers Neurorobotics*, vol. 7, p. 21, Dec. 2013.

[42] M. Janosi A. Steinbrunn, W. Pfisterer, and R. Detrano, "Heart disease," UCI Mach. Learn. Repository, 1989, doi: 10.24432/C52P4X. Accessed: May 12, 2024.

[43] N. Wolberg W. Mangasarian, O. Street, and W. Street, "Breast cancer Wisconsin (diagnostic)," UCI Mach. Learn. Repository, 1993, doi: 10.24432/C5DW2B. Accessed: May 12, 2024.

ABDULRAHMAN ALZAHRANI is currently an Associate Professor with the University of Hafr Al Batin. His research interests include computer vision, optimization techniques, and performance enhancement. His recent research interests are related to data mining, mainly working on machine learning and deep learning-based the IoT, text mining, and computer vision tasks.

• • •