EARLY DETECTION OF LUNG CANCER USING PREDICTIVE MODELING IN CORPORATING CTGAN FEATURES AND TREE-BASED LEARNING

Savitasoma[1], Veerendra Patil[2]

Access Online

Article DOI:
10.5281/zenodo.21374616

[1] Research Professor, Department of Electronics and Communication Engineering, Guru Nanak Dev Engineering College, Bidar, Karnataka, India
[2] PG Student, Department of Electronics and Communication Engineering, Guru Nanak Dev Engineering College, Bidar, Karnataka, India

How to Cite:
Savitasoma & Veerendra Patil (2026), Early Detection of Lung Cancer Using Predictive Modeling incorporatingctganfeaturesandtree-Based Learning, International Educational Journal of Science & Engineering (IEJSE),Vol: 9, Special Issue, May 2026 107-114

ABSTRACT

Lung cancer remains a deadly tumor globally, and accurate and early diagnostic methods are needed to improve the survival rate. The integration of artificial intelligence techniques with predictive modeling offers efficient techniques for the extraction of complex medical data and the support of decision-making in healthcare. The data used was obtained from Kaggle, and includes a plethora of lung cancer features extracted from patients' medical data. The dataset comprises medically relevant attributes for classification. A feature selection approach involving Chi² and MI was used to identify significant predictors, thereby supplying an effective feature set for the classification models in high-dimensional data. The classification was performed using several ML algorithmswith5-foldcross-validation -GBC, RF, ETC, SVM, KNN, LR, DT, SGDC, NB, and XGB. Various class balancing techniques, such as SMOTE, Borderline SMOTE, SMOTE-ENN and CTGAN, were applied to reduce class imbalance, thereby significantly improving model performance. To enhance accuracy, the data was preprocessed by removing outliers and using advanced resampling techniques (ADASYN+ Cluster Centroids) in over-under sampling, and GBC achieved the highest accuracy of 99.05%. Finally, to improve model interpretability, the explainable AI interpretability methods LIME and SHAP were applied, and a Flask-based application was developed to allow user input for real-time prediction.

KEYWORDS: Lung Cancer, Accuracy, Predictive Model, Data Models, Biological System Modeling, Medical Diagnostic Imaging, Support Vector Machines, Analytical Models, Prediction Algorithms, Nearest Neighbor Methods".

I. INTRODUCTION

AI has initiated a paradigm shift in many areas, such as engineering, manufacturing, and in particular, medical research. The application of ML techniques has allowed the health care industry to develop automated systems that help to collect and analyse the vast amount of medical data and support health practitioners to make accurate and effective treatment decisions. Lung cancer accounts for many cancer deaths. The symptoms in the early stages are often absent and therefore, early diagnosis and treatment are essential to improve survival rates and avoid death. Lung cancer was the most burdensome cancer in 2017 with the most DALYs and affected 40. 9 million people worldwide. In 2020,1.8million people died of lung cancer and in 2022, there will be 135,720 deaths from lung cancer in the US [1], [2]. However, lung cancer is often a symptomatic in the early stages making early detection difficult.

In the current study, we are discussing the detection and evaluation of the incidence and progression of lung cancer using ML and DL models [3]. Smoking is the major risk factor. SCLC (10-16% of cases) develops rapidly and spreads easily. ML method should be used to identify new patterns and associations in a large data set for early detection of lung cancer.

Research Paper

Recent advancements in bioinformatics, covering image-based and factor-based data, have allowed the identification of relevant features for prediction. These factors, when combined with powerful ML models, enhance the accuracy of disease predictions [4]. Various techniques including LR, RF, DT, SVM, NB and KNN have been applied to gene expression data for predicting cancer models. There are ADB-based prediction models developed to help physicians in ECOG PS-based clinical decision-making for lung cancer patients [5].

Recent studies emphasise the importance of model validation to predict cancer, including the confusion matrix, precision,recall,accuracyandF1-score[6], [7]. Sophisticated classifiers, such as XG Boost, ETC, SSGDC, GBC and ensemble models, have shown promise in early diagnosis of lung cancer from clinical and medical data [8].Furthermore, new approaches have been explored such as the use of handcrafted electronic nose (e-nose) devices for lung cancer diagnosis[9],[10].Despite such achievements, there remains much room for improvement in classification models, prediction accuracy, and the use of disparate data sources for increased clinical relevance.

II. RELATED WORK

There has been significant attention to the use of ML and artificial intelligence for medical diagnosis, particularly for complex diseases such as lung cancer. Recent studies have explored a range of approaches, datasets and models to improve early diagnosis, risk assessment and therapeutic options. Dirik [11] has developed an ML approach for diagnosing lung cancer, which showcased the power of classification algorithms that can analyse medical data for lung cancer patterns. The purpose of this study was to improve diagnosis accuracy with the help of computer models that are more reliable and repeatable than clinical decision making. Mohan and Thayil [12] have discussed the application of ML algorithms for lung cancer on the text data. They noted that clinical text data might be a treasure-trove for feature selection and classification, and that non-imaging data could be a potentially important feature in predictive models.

Fatokietal. [13] studied the application of ML techniques for lung cancer prediction using various data sets. They explored several approaches and highlighted the importance of model selection to achieve reliable results. The findings indicated ensemble learning algorithms perform better than single classifiers and pointed out hybrid approaches might be beneficial in medical diagnosis. Similarly, Tomassini Tal.

[14] conducted a detailed evaluation of lung nodule classification and cancer histological classification using CNNs based on CT images. They found CNNs have a good ability to extract spatial features in CT scans, and as a result, considerably enhance classification accuracy. The study provided a detailed summary of different CNN models and their uses, highlighting the importance of DL to achieve accurate histological classification.

Dritsas and Trigka [15] explored lung cancer risk prediction by creating and evaluating a range of ML models and discussed how statistical and computational techniques can be used to help identify the most at-risk people from clinical features. They highlighted ML models to deal with imbalanced data, while exhibiting the differences in prediction accuracy between different models. Kumar et al. [16] took a different approach and worked on the text dataset for lung cancer risk prediction, using ML algorithms to identify semantic and medical phrases. This demonstrated that natural language processing and ML can be used to develop effective diagnostic tools. Vasudha Rani et al. [17] developed lung cancer risk prediction model using ML. Their approach considered the importance of feature selection and data preparation to improve input data sets, thus resulting in better predictive outcomes. This research highlighted the importance of data preprocessing and tuning in ML for medical data.

Gültepe [18] undertook a comparative study of classification algorithms, comparing the performance of a number of lung cancer prediction techniques using various algorithms. He confirmed that algorithm choice has a big impact on accuracy, recall and precision, suggesting that no one technique always outperforms the others, but is heavily dependent on the characteristics of the dataset. Guoetal.[19] proposed a new method by advocating targeted

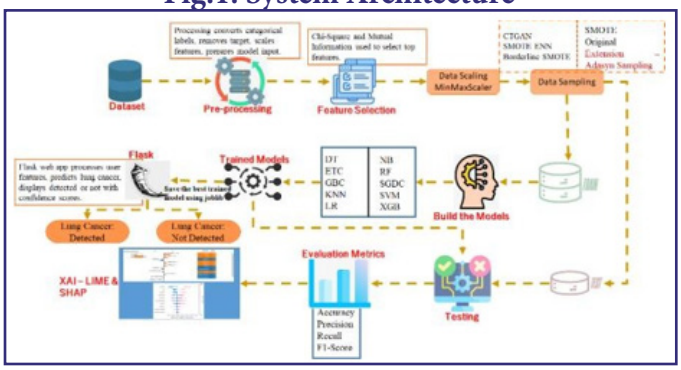
therapy to lung cancer sufferers with malignant pleural effusion using autologous tumor cell-derived microparticles. The biomedical discovery, as opposed to just computational approaches, integrated the concept of targeted therapy with clinical practice and thus linked predictions for diagnosis with targeted techniques. Their study showed the integration of computational predictions and treatment planning to improve patient care.

Chawla etal. [20] developed the SMOTE, a widely used method for balancing class distributions in medical data. Although not limited to lung cancer research, SMOTE has been widely used in the field of oncology to enhance the learning process of ML models, where the number of positive cancer patients is relatively small, compared to negative patients. This is particularly important for lung cancer prediction as data sets are often highly unbalanced in terms of a disproportionate number of early stage instances, thus limiting the models' generalization capabilities.

III. MATERIALS AND METHODS

Our approach provides a ML-based automated predictive modeling framework for the early detection of lung cancer using the Kaggle lung cancer dataset, which includes clinical and diagnostic data. For efficiency, feature selection is carried out by means of dimensionality reduction usingChi^2 and MI methods. To address data imbalance, techniques such as SMOTE, Borderline SMOTE, SMOTE-ENN, CTGAN, as well as a novel hybrid resampling approach combining ADASYN and Cluster Centroids are employed. Data pre-processing involves out Lier removal to improve the quality of the data. The approach is to train various ML algorithms including GBC, RF, ETC, SVM, KNN, LR, DT, SGDC, NB and XGB using 5-foldcross-validation. A hybrid LIME and SHAP for explainable AI and Flask based web interface for real-time prediction is developed.

Fig.1: System Architecture



The lung cancer screening process begins with a data set of lung cancer disease, which is pre-processed to correct the errors and transform the features. We split the data into training and testing data set by 5-fold cross-validation training. We use CTGAN to create new features to overcome class imbalance and enhance diversity. The data is then fed into a Random Forest model to get a better prediction result. We then evaluate the performance with a confusion matrix, between the predicted versus observed values.

A) Data Set Collection

In this research, a lung cancer data set has been used, of 309 patients with 16 features collected through questionnaire responses. These are called personal attributes (age and gender) and lifestyle and medical factors (smoking, alcohol consumption, anxiety and chronic disease). These are used to explore the relationships with lung cancer. The data collection offers an organized foundation for data processing, feature extraction and training for predictive models and classification of lung cancer rates.

B) Pre-Processing

Preprocessing ensures the data quality and readiness for modelling by visualization, cleaning, feature selection, scaling and balancing, which improves model training and testing for accurate lung cancer prediction models.

i) Data visualization: is an essential part for understanding the data by visualizing distributions. This includes plotting the target classes to ensure that there is no class imbalance and plotting correlation matrices to understand the relationships between clinical features. Visualizations help in finding important associations between the features and lung cancer prognosis, feature selection, and guides

decisions about data preprocessing and selection of ML model for effective prediction.

ii) Feature selection identifies the most relevant features that have a significant impact on lung cancer survival. The importance of each variable is measured and the relevant features are retained using Chi^2 and MI methods. The feature dimensionality is decreased by removing irrelevant or redundant attributes, which results in faster learning, improved accuracy and interpretability. This step helps in identifying the important features to improve classification accuracy.

iii) Data scaling ensures that during model training, the features are equally weighted as they will have the same range. Use Min Max Scaler or Standard Scaler as per algorithm. It alleviates the bias of the attributes that have a greater magnitude and it also speeds up convergence in the learning process. Scaling is required for distance based algorithms such as KNN, SVM etc. to ensure stability and efficiency and give equal weight to all the features in the data set.

iv) Data balancing addresses the class imbalance issue, a common issue in medical data sets because of the imbalance between positive and negative examples. We use up-to-date resampling methods, like SMOTE, Borderline-SMOTE, SMOTE-ENN and CTGAN to create instances of minority class with a balanced number of classes. This approach makes the model more objective and less biased towards majority classes and offers more accurate and stable prediction in lung cancer diagnosis.

C) Training & Testing

Training and Testing The data is divided into two parts , generally 80/20 ratio, in training and testing Training provides data to train ML algorithms and testing verifies the algorithms. This step measures the performance of the algorithms and then guarantees the predictions' correctness in real-life. The results with balanced and unbalanced data also demonstrate the effect of pre-processing and balancing techniques on the improvement of lung cancer prediction.

D) Integration of XAI and Flask Framework:

The incorporation of XAI in ML ensures interpretability of lung cancer prediction. We used techniques such as LIME and SHAP to provide explainability of model predictions (local and global) by showing feature importance for a given prediction and general trends. This instills trust in the predictions made by medical experts and makes the system trustworthy for clinical purposes. The explanations are Visualized by waterfall and summary plot graphs.

For better user experience, a web application using Flask technology was developed to ensure efficient use of the model by doctors and users. The technology allows for the input of clinical features and output of predictions and explanations. The integration of predictive models, explainable AI (XAI) and Flask technology for deployment improve the effectiveness of technology and its user experience to provide the right diagnosis and explainable medical decisions.

E) Algorithms

XG Boost (XGB): XG Boost are weighted loss functions with boosted (sequential) decision trees to assess non-linear relationships between clinical variables. Regularization prevents overfitting, hyperparameters and balanced data for precise, fast and consistent classification of lung cancer.

$$\hat{y}_i = (\sum_{k=1}^{K} f_k(x_i)),\ f_k \in F \quad (1)$$

Extreme Trees Classifier (ETC): ETC creates random forests with random splits. ETC reduces variance by averaging. A light pre-processing of diagnostic features leads to improved generalization, scalability and interpretability of feature interactions, leading to good accuracy in classification if minority classes are boosted by synthetic balancing techniques.

Support Vector Machine (SVM): SVM is used to discover the best hyperplane to separate cancer and non-cancer patients. It uses kernel functions for non-linear boundaries and balancing of classes to detect minor classes. It maximises the margin and avoids overfitting for better prediction on clinical data.

$$minimize \frac{1}{2}||W||^2 + C\sum_{i=1}^{n} \xi_i \quad (2)$$

K-Nearest Neighbors (KNN): KNN measures closeness of patients by a vote of the local neighbour

hood of feature points. It identifies small clinical patterns in balanced data with normalised features. Its simplicity and versatility offer good prediction accuracy, recall and F1-scores for cancer diagnosis.

$$distanc(x,X_i)=\sqrt{\sum_{j=1}^{d}(x_j-X_{i_j})^2} \quad (3)$$

Stochastic Gradient Descent Classifier (SGDC): SGDC learns from large data sets by repeatedly adjusting the weights to maximise the loss functions. Scaling and re-sampling for fairness for courses. Choice of loss and regularization for tuning for accuracy or recall.

Logistic Regression (LR): LR uses a logistic function to model the cancer risk and gives useful parameters to estimate the effect sizes. Random selection and regular inaction improve the stability for multicollinearity. Threshold tuning is a convenient approach to regulate complex ensemble processes, which can be achieved by using probabilistic predictions.

$$\hat{y}_i=\sigma(w^Tx+b)\equiv \frac{}{1+e^{-(w^Tx+b)}} \quad (4)$$

Decision Tree (DT): DT is a rule-based classic cation technique, which divides features into multiple thresholds. It can be applied to discrete or continuous data. It can be applied to diagnostic features of interest. In medical applications, resampling is liable to move data. but enhances generalisation and understanding of the model through visualisation.

$$(i)= 1-\sum_{i=1}^{k} p^2_i \quad (5)$$

Random Forest (RF): RF is more robust and reliable by combining multiple decision trees to get her via resampling. It can handle complex relationships and high dimensionality of medical data. Here, balanced resampling favours the minority class and key features of cancer in order of importance.

$$Gini= 1-\sum_{i=1}^{C}(Pi)^2 \quad (6)$$

Gradient Boosting Classifier (GBC): GBC grows stress to correct the errors of previous trees and optimally captures complex medical features. It uses synthetic balancing for imbalanced data. The model's accuracy, probabilistic analysis and robust early cancer diagnosis is enhanced by carefully tuning the learning rate and tree depth to avoid overfitting.

Naïve Bayes (NB): NB apply Bayes' theorem and conditional independence to assess the likelihood of cancer. The model is fast and yields probabilities. Combined with balanced data, it offers reliable initial predictions and serves as a minimal baseline model for complex systems.

$$P(A|B)= \frac{P(B|A).P(A)}{P(B)} \quad (7)$$

IV. EXPERIMENTAL RESULTS

Accuracy: The accuracy of a test is the ability of a test to correctly identify diseased and normal cases. To determine the accuracy of a test, one needs to calculate the proportion of true positives and true negatives out of all cases tested. This can be formulated as:

$$Accuracy= \frac{TP+TN}{TP+FP+TN+FN} (8)$$

Precision: Precision is a statistic that measures the accuracy of the cases identified by the test. Therefore, the accuracy equation is given as:

$$Precision= \frac{\text{True Positive}}{\text{True Positive+ False Positive}} (9)$$

Recall: Recall is a measure of the ability of a ML model to find all the relevant cases of a particular class. It measures the number of true positives correctly identified by a model as a proportion of the total number of positives, providing insight into a model's ability to identify instances of a particular class.

$$Recall= \frac{TP}{TP+FN} \quad (10)$$

F1-Score: The F1 score is a ML model's accuracy statistic. It combines the accuracy and recalls of a model. The accuracy statistic defines the number of correct predictions made by the model across the entire model.

$$F1Score=2*\frac{RecallXPrecision}{Recall+Precision}*100(11)$$

Table 1: PerformanceEvaluation

ML Model	Accuracy	Precision	Recall	F1-score
XGB	0.980952	0.981699	0.980952	0.980970
ETC	0.971429	0.971636	0.971429	0.971444
SVM	0.914286	0.921164	0.914286	0.913510
KNN	0.933333	0.934512	0.933333	0.933150
SGDC	0.933333	0.936935	0.933333	0.932966
LR	0.876190	0.876170	0.876190	0.876100
DT	0.971429	0.971636	0.971429	0.971444
RF	0.961905	0.961905	0.961905	0.961905
NB	0.838095	0.839816	0.838095	0.838272
GBC	0.990476	0.990667	0.990476	0.990481

The performance evaluation of ML models is presented in Table1using accuracy, precision, recall and F1-score. The Gradient Boosting Classifier attained the highest performance in all measures, asserting its superiority in lung cancer prediction.

Table 2: Performance Evaluation Table–DifferentDataBalancingTechniques

Model	CTGAN	SMOTE ENN	Borderline SMOTE	SMOTE	Original	Extension
DT	84.3 %	95.65 %	94.44 %	94.44 %	85.48 %	97.14 %
ETC	85.2 %	95.65 %	94.44 %	94.44 %	85.48 %	97.14 %
GBC	79.6 %	95.65 %	89.81 %	95.37 %	85.48 %	99.05 %
KNN	78.7 %	100.0 %	87.96 %	95.37 %	87.1 %	93.33 %
LR	72.2 %	63.77 %	88.89 %	92.59 %	87.1 %	87.62 %
NB	73.1 %	98.55 %	87.96 %	85.19 %	87.1 %	83.81 %
RF	85.2 %	95.65 %	94.44 %	94.44 %	87.1 %	96.19 %
SGDC	69.4 %	63.77 %	91.67 %	92.59 %	88.71 %	93.33 %
SVM	75.9 %	95.65 %	87.96 %	91.67 %	87.1 %	91.43 %
XGB	85.2 %	98.55 %	90.74 %	95.37 %	88.71 %	98.1%

Table (2) shows the performance evaluation of the Gradient Boosting Classifier, reaching the highest accuracy of 99.05% using the Extension method, which indicates its superior performance in predicting lung cancer compared to other oversampling methods.

Fig. 3: ComparisonGraph

In the comparison graph (Fig. 3), the algorithms are compared based on accuracy (green), precision (light green), recall (blue) and F1-score (red). XGB delivers the best performance in all aspects, indicating its higher efficacy.

Fig. 4: ComparisonGraph–DifferentDataBalancingTechniques

In Figure 4, the comparison graph shows that the maximum accuracy with the GBC was obtained using the Extension approach, compared to other oversampling methods.

Fig. 5: Ente rInput Data

Figure 5 shows the user interface for entering patient

details related to lung cancer for model prediction.

Fig. 6: Predicted Results

Figure6showsthesystem'spredictionwiththemessage: 'Lung cancer: NOT DETECTED.'

Fig. 7: Enter Input Data

Figure 7 shows the user interface for the input of patient characteristics, including the clinical and diagnostic features.

Fig. 8: Predicted Results

Figure 8 shows a screenshot of the system's prediction, with the message: 'Lung Cancer: DETECTED'.

V. CONCLUSION

The development of a predictive model for the early detection of lung cancer shows the success of ML and artificial intelligence in clinical decision making. Data from the Kaggle lung cancer data set, with comprehensive diagnostic and clinical features, was used to train and evaluate models. Chi² and MI were applied to feature selection for the identification of relevant features, thus improving the data quality and reducing dimensionality. Class imbalance, a common problem in medical data, was addressed through advanced balancing methods such as SMOTE, Borderline SMOTE, SMOTE-ENN and CTGAN thereby ensuring that the minority classes were well represented. A range of algorithms were used with 5-fold cross-validation, including GBC, RF, ETC, SVM, LR, KNN, DT, SGDC, NB and XGB. The results showed that the Gradient Boosting Classifier, when combined with preprocessing and oversampling-under sampling, achieved the highest predicted accuracy of 99.05%. Moreover, LIME and SHAP were used to interpret the results, and a Flask app was created for intuitive predictions.

Future work can include the use of DL models such as convolutional and recurrent neural networks to detect complex patterns in lung cancer data, and hence improve accuracy. We can fuse data such as images and genetic material for a more comprehensive diagnosis. Immediate implementation on the cloud can enable better access for practitioners. Improvements in explainable AI approaches will enhance transparency and build trust among users. Adaptive learning can be used to update models with new data, to provide flexibility. Electronic health record and mobile health systems can be integrated to enable personalised patient treatment andearly detection strategies.

REFERENCES

1. Ansari, M. M., Kumar, S., Chola, C., Heyat, M. B. B., Akhtar, F., Hayat, M. A. B., & Pomary, D. (2025). A Novel Machine and Deep Learning–Based Ensemble Techniques for Automatic Lung Cancer Detection. BioMed Research International, 2025(1), 6666688.
2. Patil, K., Dholakiya, N., Padhiyar, D., Anjaria, B., &Rana, K. (2025). A Hybrid Explainable AI Framework for Early Lung Cancer Detection Using CTGAN-Augmented Clinical Data, Gene Biomarkers, and Transformer-CNN Networks. Metallurgical and Materials Engineering, 31(4), 373-379.
3. Pavithran, M. S., & Saranyaraj, D. (2025). Lung cancer risk prediction using augmented machine learning pipelines with explainable AI. Frontiers in Artificial Intelligence, 8, 1602775.
4. Pavithran, M. S., & Saranyaraj, D. (2025). Lung cancer risk prediction using augmented machine learning pipelines

with explainable AI. Frontiers in Artificial Intelligence, 8, 1602775.
5. Ramkumar, M. O., Monica, R., Jaya Kumar, D., Rajmohan, R., & Arul Selvam, V. (2025, June). Lung Cancer Detectionusing3DU-NetBased DENSENET with Multimodal Images CT scans and his to pathological Images. In 2025 International Conference on Emerging Technologies in Engineering Applications (ICETEA)(pp.1-6). IEEE.
6. 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.
7. 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.
8. 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.
9. 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.
10. Hugging Face Datasets. (2024). Lung Cancer Dataset. Accessed: May16, 2024. [Online]. Available: https://huggingface.co/datasets/nateraw/lungcancer
11. M. Dirik, "Machine learning-based lung cancer diagnosis," Turkish J. Eng., vol. 7, no. 4, pp. 322–330, Oct. 2023.
12. K. Mohan and B. Thayyil, "Machine learning techniques for lung cancer risk prediction using text dataset," Int. J. Data In format. Intell. Comput., vol. 2, no. 3, pp. 47–56, Sep. 2023.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. Y. Gültepe, "Performance of lung cancer prediction methods using different classification algorithms," Comput., Mater. Continua, vol.67, no. 2, pp. 2015–2028, 2021.
19. 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–derivedmicroparticlebasedtargetedchemotherapyinlung-cancerpatientswithmalignant pleural effusion," Sci. Transl. Med., vol. 11, no. 474, Jan.2019, Art. no. eaat5690.
20. 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.
21. A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the15-year anniversary," J. Artif. Intell. Res., vol. 61, pp. 863–905, Apr.2018.
22. 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. In format., vol. 19, no. 3, pp. 3208–3218, Mar. 2023.
23. 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.
24. L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Proc. Adv. NeuralInf. Process. Syst., vol. 32, 2019, pp. 7335–7345.
25. D. Unzueta. (2021). How to Generate Tabular Data Using CTGANS. Towards Data Science. Accessed: 31stJan.2025. [Online].Available:https://towardsdatascience.com/how-to-generate-tabular-data-using-ctgans-d91a56054955