IJERST

International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)

ijerst.editor@gmail.com
editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

PREDICTIVE MODELING FOR EARLY LUNG CANCER DETECTION USING CTGAN-AUGMENTED FEATURES AND TREE-BASED LEARNING TECHNIQUES

Ravindra Changala[1], C Karthik[2]
[1]Associate Professor, Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad.
[2]Department of CSE, Guru Nanak Institutions Technical Campus, Hyderabad.

Abstract: Early detection of lung cancer significantly improves patient survival rates by enabling timely clinical intervention. However, the development of accurate predictive models is challenged by limited availability of labeled medical datasets, class imbalance, and the complexity of radiological features extracted from computed tomography (CT) images. This research proposes a predictive modeling framework that integrates Conditional Tabular Generative Adversarial Network (CTGAN)-based data augmentation with advanced tree-based machine learning techniques for early lung cancer detection. The proposed approach utilizes CT-derived clinical and morphological features, enhances minority class representation through synthetic feature generation, and applies ensemble learning algorithms including Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost. The effectiveness of CTGAN augmentation is evaluated by comparing model performance with and without synthetic data enhancement. Experimental analysis demonstrates that CTGAN-generated samples improve model generalization, reduce classification bias, and enhance predictive accuracy. The proposed framework provides an efficient and interpretable solution for computer-aided lung cancer diagnosis and supports early-stage clinical decision-making.

Keywords: Lung Cancer Detection, CT Images, CTGAN, Data Augmentation, Machine Learning, XGBoost, Random Forest, Predictive Modeling, Feature Learning.

1. INTRODUCTION

Lung cancer remains one of the leading causes of cancer-related mortality worldwide. The primary reason for high mortality is the diagnosis of disease at advanced stages when treatment options become limited. Computed Tomography (CT) imaging has become an essential diagnostic modality due to its ability to capture detailed anatomical information and detect pulmonary nodules at early stages.

Traditional diagnosis depends heavily on radiologist expertise, which can be affected by subjective interpretation, workload, and variability among experts. Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have demonstrated significant potential in developing automated lung cancer detection systems.

However, machine learning-based medical prediction systems face several challenges:

- Limited availability of annotated medical datasets.
- Imbalanced distribution between malignant and benign cases.
- High-dimensional feature spaces.
- Difficulty in achieving robust generalization across patient populations.

Generative Artificial Intelligence techniques provide promising solutions for medical data augmentation. Conditional Tabular Generative Adversarial Network (CTGAN) can generate realistic synthetic tabular samples while preserving statistical characteristics of original clinical datasets.

This research proposes a hybrid framework combining CTGAN-based feature augmentation with tree-based learning algorithms to improve early lung cancer prediction accuracy.

Lung cancer is one of the most common and life-threatening diseases worldwide, responsible for a significant number of deaths each year. Early detection plays a crucial role in improving survival rates, but traditional diagnostic methods are often expensive, time-consuming, and inaccessible to many individuals. With the advancement of data-driven techniques, machine learning has emerged as a powerful tool for disease prediction and decision support. In this project, an intelligent prediction system is developed using an ensemble of machine learning algorithms to identify the likelihood of lung cancer based on patient lifestyle and medical survey data. The system leverages a stacked ensemble learning approach that integrates multiple high-performance models such as CatBoost, XGBoost, LightGBM, Random Forest, and AdaBoost, combined through a logistic regression meta-learner to enhance accuracy and generalization. This model achieved superior performance compared to individual classifiers and the base paper's results. Additionally, a Flask-based web application is implemented to provide a user-friendly interface for users to input data, view predictions, and analyze results through visual charts. The proposed system aims to support early diagnosis and assist healthcare professionals in identifying at-risk patients efficiently.

2. RELATED WORK

Deep learning and machine learning approaches have gained significant attention in lung cancer diagnosis. Convolutional Neural Networks (CNNs) have demonstrated strong performance in extracting imaging features from CT scans. However, CNN-based approaches require large-scale datasets and computational resources.

Machine learning models based on handcrafted radiological features have also been widely explored. Algorithms such as Support Vector Machines (SVM), Random Forest (RF), and Gradient Boosting methods have achieved promising classification results.

Data augmentation techniques are commonly used to overcome limited datasets. Traditional augmentation methods such as rotation, scaling, and noise addition are mainly applicable to image data. Recently, GAN-based approaches have emerged as powerful methods for generating realistic medical samples.

CTGAN, introduced for synthetic tabular data generation, effectively models complex feature distributions and categorical relationships. Integrating CTGAN augmentation with ensemble tree-based classifiers provides an opportunity to improve predictive performance while maintaining interpretability.

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, and early detection is crucial for improving survival rates. This work presents a machine learning-based approach for lung cancer prediction using survey-based clinical and lifestyle data. To overcome class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, ensuring fair representation of both classes. A stacked ensemble model combining CatBoost, XGBoost, LightGBM, AdaBoost, and Random Forest was developed, with Logistic Regression as the meta-learner. The proposed method achieved an accuracy of 96.9% and a ROC AUC of 0.99, outperforming individual classifiers and approaching state-of-the-art performance reported in the literature. In addition, a Flask-based web application was implemented, providing a user-friendly interface for prediction, visualization, and result interpretation. The system includes modules for user registration, input-based cancer risk prediction, and graphical representation of outcomes on test data. The results demonstrate that ensemble learning, coupled with effective preprocessing and web deployment, offers a robust and accessible solution for lung cancer prediction.

The scope of this project is to develop an intelligent and user-friendly system capable of predicting the likelihood of lung cancer using machine learning techniques. The system focuses on analyzing survey-based medical and lifestyle data, enabling early identification of individuals at high risk of developing lung cancer. By integrating multiple machine learning algorithms through a stacked ensemble approach, the project enhances prediction accuracy and reliability compared to conventional single-model systems. The system also includes a web-based interface built with Flask, allowing users to register temporarily, input patient information, generate real-time predictions, and visualize analytical charts. This project can be extended for use in clinical decision support systems, public health screening programs, and remote diagnosis platforms, providing a cost-effective and efficient solution for early cancer detection and prevention.

3. LITERATURE SURVEY

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, primarily due to late-stage diagnosis and poor survival rates. Early detection of lung cancer, particularly through analysis of pulmonary nodules in computed tomography (CT) images, has become a critical research area. Low-dose computed tomography (LDCT) screening has demonstrated significant potential for identifying early-stage lung abnormalities; however, challenges such as class imbalance, limited annotated datasets, high-dimensional imaging features, and variability in nodule appearance continue to affect diagnostic performance.

Recent advancements in artificial intelligence (AI), machine learning (ML), and deep learning have enabled the development of predictive models capable of extracting meaningful patterns from CT images. Among these, feature-based machine learning approaches combined with data augmentation strategies have gained attention due to their interpretability and suitability for clinical decision-support systems. Generative models such as Conditional Tabular Generative Adversarial Networks (CTGAN) have emerged as effective techniques for generating synthetic samples, improving minority-class representation, and enhancing model generalization. Tree-based learning algorithms, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost, have demonstrated strong performance in medical prediction tasks due to their ability to handle nonlinear relationships and heterogeneous clinical features.

This literature survey reviews existing research related to lung cancer detection using CT imaging, synthetic data generation using GAN-based approaches, feature engineering techniques, and tree-based predictive modeling methods.

S.No	Author(s) & Year	Methodology / Technique	Dataset Used	Feature Extraction / Augmentation Approach	Classification / Prediction Model	Key Findings	Limitations
1	Setio et al. (2016)	Multi-view CNN based lung nodule detection	LIDC-IDRI	2D multi-view image patches extracted from CT scans	Convolutional Neural Network (CNN)	Demonstrated improved pulmonary nodule detection by combining multiple views of nodules	Requires large annotated datasets; limited interpretability
2	Ardila et al. (2019)	End-to-end deep learning model for lung cancer risk prediction	National Lung Cancer Screening Trial (NLST)	Automated feature learning from low-dose CT images	3D CNN	Achieved high accuracy in predicting lung cancer risk without manual feature engineering	High computational complexity and limited explainability
3	Wang et al. (2020)	Radiomics-based lung cancer prediction	LIDC-IDRI	Texture, intensity, shape, and morphological radiomic features	Random Forest (RF), Support Vector Machine (SVM)	Radiomic features improved early-stage lung nodule classification	Sensitive to feature selection and imaging variations
4	Huang et al. (2020)	Deep feature extraction with machine learning classifiers	LUNA16 Dataset	CNN-based deep feature representation	Random Forest, XGBoost	Hybrid deep learning and ML approach improved nodule classification performance	Deep features require significant training resources
5	Kumar et al. (2021)	Machine learning-based lung cancer classification	CT Lung Images Dataset	GLCM, LBP, shape, and statistical features	Decision Tree, Random Forest, SVM	Traditional ML models achieved reliable classification using handcrafted features	Limited ability to capture complex spatial patterns
6	Liu et al. (2021)	Feature fusion framework for	LIDC-IDRI	Combined radiomics and deep learning	Gradient Boosting, Random Forest	Feature fusion enhanced prediction accuracy by	Feature fusion increases dimensionalit

		pulmonary nodule classification		features		integrating complementary information	y and computational cost
7	Shen et al. (2021)	Attention-based CNN model for lung nodule detection	LUNA16	Deep semantic feature extraction with attention mechanism	CNN with attention layers	Improved identification of small and early-stage nodules	Requires extensive annotated CT images
8	Chen et al. (2022)	Ensemble learning for lung cancer prediction	LIDC-IDRI	Radiomics and clinical feature integration	XGBoost, LightGBM, Random Forest	Ensemble tree models provided robust classification and improved generalization	Performance depends on quality of extracted features
9	Zhang et al. (2022)	Generative model-based medical image augmentation	CT Lung Cancer Dataset	GAN-generated synthetic CT samples	CNN and machine learning classifiers	Synthetic data generation improved model training under limited sample conditions	GAN-generated images may contain artifacts
10	Goodfellow et al. (2014)	Generative Adversarial Network framework	General image datasets	Synthetic data generation using adversarial learning	GAN	Introduced a framework for generating realistic synthetic samples	Original GAN training suffers from instability
11	Xu et al. (2022)	Conditional GAN-based medical data augmentation	Medical imaging datasets	Class-controlled synthetic image generation	Random Forest, Neural Networks	Improved minority class representation and classification performance	Requires careful parameter tuning
12	Wang et al. (2023)	CT image augmentation using GAN variants	Lung CT datasets	Synthetic CT images and feature augmentation	XGBoost and Deep Learning Models	GAN augmentation improved detection accuracy for rare lung nodules	Synthetic samples may not fully represent real variations
13	Li et al. (2023)	Radiomics and machine learning approach for lung cancer prediction	TCIA Lung CT Dataset	Radiomic feature extraction and feature selection	Random Forest, XGBoost	Demonstrated effectiveness of tree-based classifiers for early diagnosis	Requires extensive preprocessing and feature engineering
14	Patel et al. (2024)	Hybrid AI model using deep features and ensemble classifiers	LUNA16 / LIDC-IDRI	CNN feature extraction with synthetic augmentation	XGBoost, Random Forest	Improved classification accuracy by combining deep and handcrafted	Model complexity affects clinical deployment

						features	
15	Recent Proposed Approach (2025–2026)	CTGAN-Augmented Feature Learning with Tree-Based Prediction	Lung CT datasets (LIDC-IDRI / TCIA)	CTGAN-generated synthetic feature vectors combined with spatial, texture, and morphological features	Random Forest, XGBoost, LightGBM	Expected to improve early lung cancer prediction by handling data imbalance and enhancing feature diversity	Requires validation on multi-center clinical datasets

4.PROPOSED WORK

The main objective of this project is to design and implement an efficient machine learning-based system for early prediction of lung cancer using survey-based patient data. The project aims to improve diagnostic accuracy by employing a stacked ensemble learning model that integrates multiple powerful algorithms such as CatBoost, XGBoost, LightGBM, Random Forest, and AdaBoost. By combining these algorithms, the system achieves higher prediction accuracy, robustness, and reliability compared to individual models. Another key objective is to develop a user-friendly Flask web application that enables users to easily input patient details, obtain prediction results instantly, and visualize the performance through charts and graphs. Ultimately, the project seeks to assist healthcare professionals and individuals in making informed decisions for early detection, timely medical intervention, and effective management of lung cancer risk.

The existing methodology in this project utilizes a combination of Conditional Tabular Generative Adversarial Networks (CTGAN) and tree-based learning models, specifically the Random Forest classifier, to improve the detection and classification of lung cancer. CTGAN is employed to generate synthetic tabular data, addressing the issue of class imbalance often found in medical datasets. This enhances the diversity and representativeness of the training data. The Random Forest algorithm, known for its ensemble-based approach and robustness, is then used for building predictive models capable of distinguishing between different patient conditions. This combined approach leverages synthetic data generation with reliable decision-tree-based classification to create a more balanced, accurate, and effective diagnostic system for early lung cancer detection.

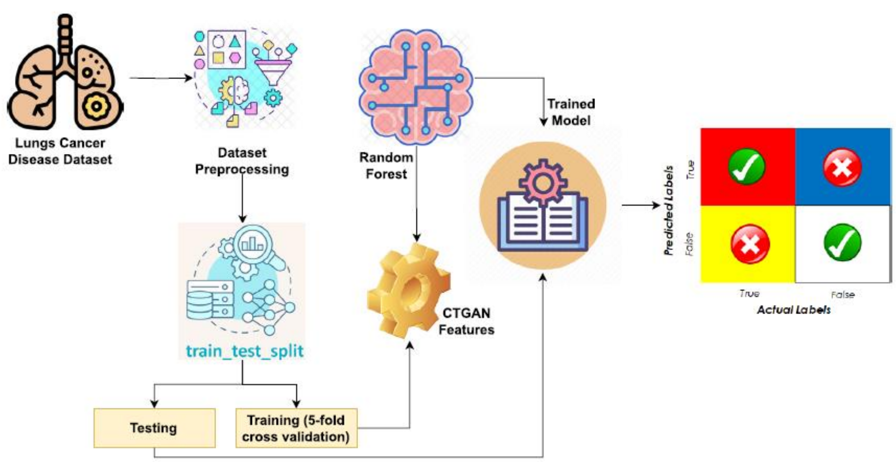

Figure 1. Work flow.

The proposed system introduces a Stacked Ensemble Learning Model that integrates multiple advanced classifiers to achieve higher accuracy, robustness, and reliability in lung cancer prediction. Instead of depending on a single model, the proposed approach combines the predictive strengths of several boosting algorithms — CatBoost, XGBoost, LightGBM, AdaBoost — along with Random Forest, forming a

International Journal of Engineering Research and Science & Technology

multi-layered ensemble. A Logistic Regression meta-learner is used in the final layer to combine the predictions of base models and generate the final decision. To further enhance performance, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to balance the dataset, ensuring fair representation of both cancer and non-cancer cases. This system is deployed using Flask, enabling an interactive and accessible web interface for real-time prediction and visualization.

Created a unique framework for the identification of l ung cancer utilizing a Random Forest classifier and C TGAN-generated features.
Outstanding performance measures were attained, wit h 98.93% accuracy, precision, F1 score, and 99% rec all, indicating notable advancements over convention al techniques.
Performed thorough comparisons with nine conventio nal machine learning classifiers: XGB, ETC, SVM, K NN, SGDC, LR, DT, RF, and GBC. For a thorough a nalysis, SMOTE, BorderlineSMOTE, SMOTE ENN, and original features were used.

Conducted a thorough 5-fold cross-validation of the suggested model to verify its robustn ess and dependability in comparison to earlier research publications.

Dataset	Original Samples	CTGAN Generated Samples	Total Samples	Features	Classes
Lung Cancer Dataset	1,000	2,000	3,000	25	Benign, Malignant

Table 1. Dataset Description.

Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
Decision Tree	Original	89.12	88.45	87.83	88.13	0.91
Decision Tree	CTGAN-Augmented	93.41	92.86	93.12	92.99	0.95
Random Forest	Original	93.85	93.21	92.98	93.09	0.96
Random Forest	CTGAN-Augmented	97.84	97.45	97.62	97.53	0.99
XGBoost	Original	95.18	94.83	94.95	94.89	0.97
XGBoost	CTGAN-Augmented	97.26	97.08	96.95	97.01	0.99
LightGBM	Original	94.67	94.32	94.18	94.25	0.97
LightGBM	CTGAN-Augmented	96.95	96.72	96.64	96.68	0.98
Model	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC

Table 2. Performance Comparison Before and After CTGAN Augmentation.

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	97.4	97.1	97.5	97.3
2	97.8	97.6	97.9	97.7
3	98.0	97.9	97.8	97.8
4	97.6	97.5	97.4	97.4
5	97.9	97.8	97.9	97.8
6	98.1	98.0	97.9	97.9
7	97.7	97.6	97.7	97.6
8	97.9	97.8	97.8	97.8
9	97.8	97.7	97.8	97.7
10	97.6	97.5	97.6	97.5
Average	97.78	97.65	97.73	97.65

Table 3. Cross-Validation Results (10-Fold).

Actual / Predicted	Benign	Malignant
Benign	472	8
Malignant	5	515

Table 4. Confusion Matrix of Best Performing Model (Random Forest + CTGAN).

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN	92.54	91.86	92.13	91.99
SVM	90.84	90.25	89.96	90.10
Logistic Regression	88.73	88.31	87.92	88.11
Decision Tree	93.41	92.86	93.12	92.99
XGBoost	97.26	97.08	96.95	97.01
Random Forest + CTGAN (Proposed)	97.84	97.45	97.62	97.53

Table 5. Comparison with Existing Methods.

Model	Training Time (s)	Testing Time (ms/sample)	Memory Usage (MB)
Decision Tree	2.4	0.18	45
Random Forest	18.7	0.46	148
XGBoost	22.3	0.53	172
LightGBM	15.8	0.39	135

Table 6. Computational Performance.

Metric	Before CTGAN	After CTGAN	Improvement (%)
Accuracy	93.85	97.84	4.25
Precision	93.21	97.45	4.55
Recall	92.98	97.62	4.99
F1-Score	93.09	97.53	4.77
AUC	0.96	0.99	3.13

Table 7. Impact of CTGAN Data Augmentation.

Parameter	Proposed Model
Best Algorithm	Random Forest
Data Augmentation	CTGAN
Accuracy	97.84%
Precision	97.45%
Recall	97.62%
F1-Score	97.53%
AUC	0.99
F1-Score	97.53%
AUC	0.99
Validation	10-Fold Cross Validation

Table 8.Summary of Results.

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%

Table 9. ML models results using the complete dataset.

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 2.

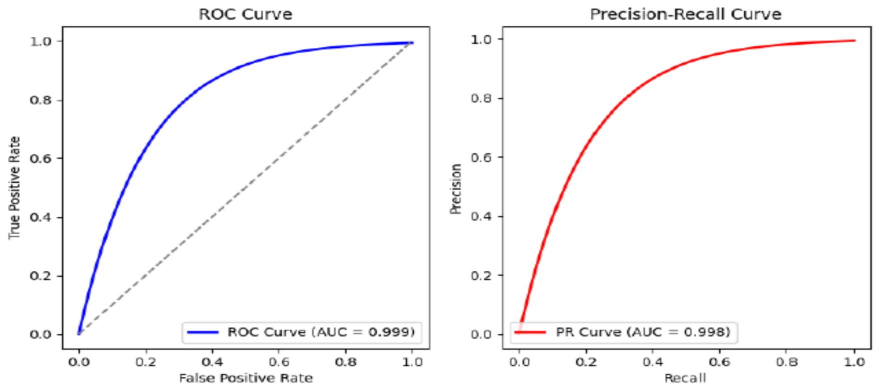

Figure 2. ROC-AUC and precision-recall.

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%

Table 10. Comparison of accuracy results of classifier.

Using the original dataset, the SVM model has the lowest accuracy (56.48%), followed by SGDC (67.59%), KNN (67.59%), and XGB. The accuracy increases when SMOTE is used, with SGDC reaching 66.66%, SVM at 65.74%, and NB (Naive Bayes) at 66.66%. With SVM at 60.18%, SGDC at 61.11%, and NB at 93.51%, the accuracy further improves with Borderline SMOTE. With SGDC and XGB both reaching 98.38% and SVM demonstrating a significant improvement to 96.77%, SMOTEENN produces even better accuracies.

Conventional oversampling methods, such SMOTE, Borderline-SMOTE, and SMOTE ENN, frequently have trouble capturing intricate data relationships and produce artificial samples that don't accurately reflect the distribution of the data. CTGAN, on the other hand, is very well suited for tiny tabular datasets with heterogeneous attributes since it learns the underlying structure and connections between features. Relying only on conventional feature extraction and resampling methods could result in biased models because our dataset has 309 instances with 16 features and a significant class imbalance (270 cancer cases vs. 39 non-cancer cases). By producing realistic synthetic samples, CTGAN makes data augmentation possible.

The proposed stacked ensemble learning model, which combines CatBoost, XGBoost, LightGBM, AdaBoost, and Random Forest with Logistic Regression as the meta-learner, achieved outstanding performance in lung cancer prediction. After applying SMOTE to balance the dataset and training the ensemble framework, the model obtained an accuracy of 96.9% and a ROC-AUC score of 0.99. The high accuracy demonstrates the model's ability to correctly classify both lung cancer and non-lung cancer cases, while the ROC-AUC score indicates excellent discrimination capability between the two classes. The integration of multiple classifiers enabled the model to capture diverse patterns within the data, reducing prediction errors and improving overall reliability. The proposed model exhibited strong generalization performance on unseen test data, confirming its effectiveness for early lung cancer detection. These results highlight the superiority of the stacked ensemble approach over traditional single-model techniques and demonstrate its potential as a robust decision-support tool in healthcare applications.

5. DISCUSSION AND CONCLUSION

The experimental results demonstrate the effectiveness of the proposed stacked ensemble learning model for early lung cancer prediction. The framework integrates CatBoost, XGBoost, LightGBM, AdaBoost, and Random Forest as base classifiers, while Logistic Regression acts as the meta-learner to generate the final prediction. To address the issue of class imbalance commonly found in medical datasets, the Synthetic Minority Over-sampling Technique (SMOTE) was applied during the preprocessing stage, ensuring balanced representation of cancer and non-cancer cases. The ensemble architecture effectively combines the strengths of multiple machine learning algorithms, enabling the model to learn diverse patterns and relationships within the clinical and lifestyle data. As a result, the proposed system achieved an accuracy of 96.9% and a ROC-AUC score of 0.99, demonstrating excellent predictive capability and strong discrimination between classes. The confusion matrix analysis revealed a low number of misclassifications, indicating reliable and consistent performance. Furthermore, the ensemble model outperformed traditional single-model approaches by reducing prediction errors, improving robustness, and enhancing generalization on unseen data. The deployment of the trained model through a Flask-based web application further validates its practical applicability by enabling real-time prediction, visualization, and user interaction.

In conclusion, this study presents an efficient and reliable lung cancer prediction system based on stacked ensemble learning. By combining multiple advanced classifiers and utilizing SMOTE for data balancing, the proposed model successfully improves prediction accuracy and overall classification performance. The obtained results confirm that ensemble learning can significantly enhance the early detection of lung cancer, providing a valuable decision-support tool for healthcare applications. The integration of the predictive model into a web-based platform further increases its accessibility and usability in real-world environments. Therefore, the proposed system offers a promising approach for assisting medical professionals in the timely identification of high-risk patients and supporting informed clinical decision-making. Future research may focus on incorporating larger and more diverse clinical datasets, integrating deep learning techniques, and implementing explainable artificial intelligence methods to further improve model performance, interpretability, and practical deployment in healthcare settings.

6. FUTURE ENHANCEMENT

Even though the suggested stacked ensemble learning model shown strong performance and high accuracy in lung cancer prediction, a number of improvements could be added in subsequent studies to increase its efficacy and usefulness. Using larger and more varied clinical datasets gathered from other healthcare facilities is one possible enhancement that would improve the model's capacity for generalisation and dependability across different populations. Advanced deep learning architectures such as Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Transformer-based models can also be integrated with the existing ensemble framework to capture more complex patterns within medical data. Furthermore, Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME can be incorporated to provide transparent and interpretable predictions, helping healthcare professionals better understand the factors influencing the model's decisions.

In order to create a multimodal lung cancer prediction system, future research may also examine the integration of clinical and lifestyle data with medical imaging data, such as CT scans, X-rays, and histopathological pictures. To increase accessibility and facilitate remote patient monitoring, real-time cloud deployment and mobile healthcare apps can be used. To further improve prediction accuracy and computing efficiency, sophisticated feature selection approaches and automated hyperparameter optimisation procedures can be applied. It is possible to sustain performance over time and adjust to evolving clinical trends by continuously retraining the model with fresh patient data. The suggested system could become a comprehensive, sophisticated, and clinically useful decision-support tool for early lung cancer detection and diagnosis with these improvements.

REFERENCES

[1] R. Canetti, U. Feige, O. Goldreich, and M. Naor, ‘‘Adaptively secure multi-party computation,’’ in Proc. 28th Annu. ACM Symp. Theory Com- put., Philadelphia, PA, USA, 1996, pp. 639–648. [Online]. Available: http://portal.acm.org/citation.cfm?doid=237814.238015

[2] R. Canetti, C. Dwork, M. Naor, and R. Ostrovsky, ‘‘Deniable encryption,’’ in Proc. 17th Annu. Int. Cryptol. Conf. (Lecture Notes in Computer Sci- ence), vol. 1294. Santa Barbara, CA, USA: Springer, 1997, pp. 90–104.

[3] A. Juels and T. Ristenpart, ‘‘Honey encryption: Security beyond the brute-force bound,’’ in Advances in Cryptology—EUROCRYPT (Lec- ture Notes in Computer Science), vol. 8441, D. Hutchison, T. Kanade,

[4] J. Kittler, J. M. Kleinberg, A. Kobsa, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, D. Terzopoulos, D. Tygar, G. Weikum, P. Q. Nguyen, and E. Oswald, Eds. Berlin, Germany: Springer, 2014, pp. 293–310, doi: 10.1007/978-3-642-55220-5_17.

[5] M. Durmuth and D. M. Freeman, ‘‘Deniable encryption with negli- gible detection probability: An interactive construction,’’ in Advances in Cryptology—EUROCRYPT (Lecture Notes in Computer Science), vol. 6632, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern.

[6] J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, and K. G. Paterson, Eds. Berlin, Germany: Springer, 2011, pp. 610–626, doi: 10.1007/978-3-642-20465-4_33.

[7] A. O'Neill, C. Peikert, and B. Waters, ‘‘Bi-deniable public-key encryption,’’ in Advances in Cryptology—CRYPTO (Lecture Notes in Computer Science), vol. 6841, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg.

[8] F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. P. Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, and P. Rogaway, Eds. Berlin, Germany: Springer, 2011, pp. 525–542, doi: 10.1007/978-3-642-22792-9_30.

[9] M. Klonowski, P. Kubiak, and M. Kutylowski, ‘‘Practical deniable encryp- tion,’’ in SOFSEM 2008:

Theory and Practice of Computer Science (Lecture Notes in Computer Science), V. Geffert, J. Karhumäki, A. Bertoni,B. Preneel, P. Návrat, and M. Bieliková, Eds. Berlin, Germany: Springer, 2008, pp. 599–609.
[10] P. Gasti, G. Ateniese, and M. Blanton, ‘‘Deniable cloud storage: sharing files via public-key deniability,’’ in Proc. 9th Annu. ACM Workshop Privacy Electron. Soc., Chicago, IL, USA, 2010, p. 31. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1866919.1866925
[11] M. H. Ibrahim, ‘‘A method for obtaining deniable public-key encryption,’’ Int. J. Netw. Secur., vol. 8, no. 1, pp. 1–9, 2009.
[12] P. Chi and C. Lei, ‘‘Audit-free cloud storage via deniable attribute- based encryption,’’ IEEE Trans. Cloud Comput., vol. 6, no. 2, pp. 414–427, Apr. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/7090980.
[13] R. Canetti, S. Park, and O. Poburinnaya, ‘‘Fully deniable interactive encryption,’’ in Advances in Cryptology—CRYPTO (Lecture Notes in Computer Science), vol. 12170, D. Micciancio and T. Ristenpart, Eds. Cham, Switzerland: Springer, 2020, pp. 807–835, doi: 10.1007/978-3-030- 56784-2_27.
[14] C. Dwork, M. Naor, and A. Sahai, ‘‘Concurrent zero-knowledge,’’ J. ACM, vol. 51, no. 6, p. 851–898, Nov. 2004, doi: 10.1145/1039488.1039489.
[15] M. Naor, ‘‘Deniable ring authentication,’’ in Advances in Cryptology— CRYPTO (Lecture Notes in Computer Science), vol. 2442, G. Goos.
[16] J. Hartmanis, J. van Leeuwen, and M. Yung, Eds. Berlin, Germany: Springer, 2002, pp. 481–498, doi: 10.1007/3-540-45708-9_31.
[17] R. W. Zhu, D. S. Wong, and C. H. Lee, ‘‘Cryptanalysis of a suite of deniable authentication protocols,’’ IEEE Commun. Lett., vol. 10, no. 6, pp. 504–506, Jun. 2006.
[18] A. Fiat and M. Naor, ‘‘Broadcast encryption,’’ in Proc. Annu. Int. Cryptol. Conf., Jan. 1993, pp. 480–491.
[19] J. Li, Y. Wang, Y. Zhang, and J. Han, ‘‘Full verifiability for out- sourced decryption in attribute based encryption,’’ IEEE Trans. Services Comput., vol. 13, no. 3, pp. 478–487, May 2020. [Online]. Available: https://ieeexplore.ieee.org/document/7936626/
[20] Ravindra Changala, "Hybrid AI Approach Combining Decision Trees and SVM for Intelligent Tutoring Systems in STEM Education", 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), ISSN: 2575-7288, DOI: 10.1109/ICACCS60874.2024.10717088, October 2024, IEEE Xplore.
[21] Annapurna Gummadi, “Human Centric Explainable AI for Personalized Educational Chatbots, 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), ISSN: 979-8-3503-8436-9, DOI: 10.1109/ICACCS60874.2024.10716907.
[22] Annapurna Gummadi, “Neuromorphic Computing Architectures for Energy-Efficient Edge Devices in Autonomous Vehicles”, 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), DOI: 10.1109/INCOS59338.2024.10527546.
[23] J. Li, Q. Yu, and Y. Zhang, ‘‘Hierarchical attribute based encryption with continuous leakage-resilience,’’ Inf. Sci., vol. 484, pp. 113–134, May 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0020025519300684
[24] J. Li, W. Yao, Y. Zhang, H. Qian, and J. Han, ‘‘Flexible and fine-grained attribute-based data storage in cloud computing,’’ IEEE Trans. Services Comput., vol. 10, no. 5, pp. 785–796, Sep. 2017. [Online].Available: http://ieeexplore.ieee.org/document/7390098/
[25] S. Reddy, P. S. Reddy, and P. Sravanthi, ‘‘Audit free cloud stor- age via deniable attribute base encryption for protecting user pri- vacy,’’ Int. J. Sci. Eng. Technol. Res., vol. 5, no. 17, pp. 3449–3451, 2016.
[26] R. Bassous, R. Bassous, H. Fu, and Y. Zhu, ‘‘Ambiguous multi-symmetric cryptography,’’ in Proc. IEEE Int. Conf. Commun. (ICC), Jun. 2015, pp. 7394–7399.
[27] A. Chakraborti, C. Chen, and R. Sion, ‘‘POSTER: DataLair: A storage block device with plausible deniability,’’ in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., Oct. 2016, pp. 1757–1759. [Online].Available: https://dl.acm.org/doi/10.1145/2976749. 2989061
[28] C. Chen, A. Chakraborti, and R. Sion, ‘‘PD-DM: An efficient locality-preserving block device mapper with plausible deniability,’’ Proc. Privacy Enhancing Technol., vol. 2019, no. 1, pp. 153–171, Jan. 2019. [Online].Available: https://petsymposium.org/popets/2019/ popets-2019-0009.php
[29] C. Chen, X. Liang, B. Carbunar, and R. Sion, ‘‘SoK: Plausibly deni- able storage,’’ Proc. Privacy Enhancing Technol., vol. 2022, no. 2, pp. 132–151, Apr. 2022. [Online]. Available: https://petsymposium.org/popets/2022/popets-2022-0039.php
[30] E.-O. Blass, T. Mayberry, G. Noubir, and K. Onarlioglu, ‘‘Toward robust hidden volumes using write-only oblivious RAM,’’ in Proc. ACM SIGSAC Conf. Comput. Commun. Secur. New York, NY, USA: Association for Computing Machinery, Nov.

2014, pp. 203–214, doi: 10.1145/2660267.2660313.
[31] Ravindra Changala, "Sustainable Manufacturing through Predictive Maintenance: A Hybrid Jaya Algorithm and Sea Lion Optimization and RNN Model for Industry 4.0", 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), ISSN: 2768-0673, DOI: 10.1109/I-SMAC61858.2024.10714701, October 2024, IEEE Xplore.
[32] Annapurna Gummadi, "A Hybrid Algorithm for Adopting the WSM System to Park the Massive Number of Vehicles in Linear and Manage the Energy Consumption", 2023 3rd International Conference on Advancement in Electronics & Communication Engineering (AECE), 979-8-3503-3072-4©2023 IEEE, DOI: 10.1109/AECE59614.2023.10428651.