Mantik Journal, 9 (1) (2025), ISSN 2685-4236 (Online)

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik

A Improving lung cancer classification with feature selection: a comparative study of random forest and xgboost

David Kurniawan[1], Ega Budiman[2], Muhammad Fadli[3], Erliyan Redy Susanto[4]
[1,2,4]Master of Computer Science, Indonesia Teknokrat University, Bandar Lampung, Lampung, Indonesia
[3]Department of Economics and Business, State Polytechnic of Lampung, Bandar Lampung, Lampung, Indonesia

A R T I C L E I N F O

Article history:

Received May 06, 2025
Revised May 19, 2025
Accepted May 30, 2025

Keywords:

Classification accuracy;
Feature selection;
Genetic algorithm;
Lung cancer classification;
Machine learning.

ABSTRACT

The leading cause of cancer mortality worldwide remains lung cancer which can be better managed when early and precise diagnosis is achieved to enhance patient outcomes. High-dimensional datasets in medical diagnostics create obstacles for classification because redundant and irrelevant features diminish model accuracy and boost computational complexity. This research investigates how feature selection enhances the performance of lung cancer classification models. The study evaluates Random Forest (RF) and XGBoost as classification models and uses Genetic Algorithm (GA) for feature selection to enhance model efficiency. The GA process ran for 50 generations and reached convergence at the 40th generation which showed that the optimal feature subset had reached stability. Random Forest outperformed XGBoost using GA-based feature selection in a number of parameters, such as accuracy, precision, recall, F1-score, and AUC-ROC. Random Forest displays superior effectiveness in utilizing optimized feature subsets to achieve enhanced generalization and classification performance over XGBoost. The research stands out because it compares how feature selection affects RF and XGBoost algorithms for lung cancer classification using fixed model settings. The research findings demonstrate the value of integrating RF with GA for feature selection as it offers potential for building both efficient and interpretable lung cancer diagnostic models within medical AI.

Corresponding Author:

Ega Budiman,
Master of Computer Science Department,
Indonesia Teknokrat University
Jl. ZA Pagar AlamNo.9-11 Labuhan Ratu Kec. Kedaton Bandar Lampung
Email: ega_budiman@teknokrat.ac.id

1. INTRODUCTION

In machine learning, feature selection is an essential step that improves model performance by removing redundant or unnecessary information while keeping the most useful ones. In high-dimensional datasets, such those in the financial and medical fields, where extraneous factors might introduce noise and reduce accuracy, this procedure is very crucial (Göltepe, 2021)(Al-Rajab, Lu and Xu, 2021). Models can generalize more

successfully and run with lower computational costs by eliminating irrelevant information, which is crucial for applications that need to process and make decisions in real time (Göltepe, 2021)(Kaur and Kumari, 2022).

Numerous feature selection strategies, such as filter-based, wrapper-based, and embedding approaches, have been developed in order to accomplish these advantages (Sheth, Patil and Dhore, 2022)(Zhang *et al.*, 2023). Moreover, metaheuristic methods such as Particle Swarm Optimization (PSO) and Whale Optimization Algorithm (WOA) have been widely employed for feature selection due to their ability to efficiently search high-dimensional feature spaces (Shami *et al.*, 2022)(Nadimi-shahraki, Zamani and Mirjalili, 2022). Additionally, recent research has shown that feature selection techniques that combine statistical and machine learning-based methodologies can improve classification performance, especially when it comes to distinguishing between subtypes of lung cancer (Chen and Dhahbi, 2021).

Filter-based methods utilize statistical tests to rank feature importance independently of any learning algorithm, making them computationally efficient (Vijayalakshmi *et al.*, 2020). In contrast, wrapper methods assess subsets of features based on model performance, often leading to better results but at a higher computational cost (Sheth, Patil and Dhore, 2022). Feature selection is carried out during training by embedded techniques like Random Forest (RF) and XGBoost, which provide a balance between efficiency and performance (Zhang *et al.*, 2023)(Khanna, Kumar and Bhat, 2025). However, metaheuristic approaches like Binary Dragonfly Algorithm (BDA) have been explored as an alternative feature selection method, showing competitive performance in optimizing feature subsets for classification tasks (Too and Mirjalili, 2021). By using semi-random feature selection, Random Forest is well known for its capacity to manage a vast number of features and determine which are most crucial (Benghazouani, Nouh and Zakrani, 2024). Meanwhile, XGBoost, on the other hand, is a scalable tree boosting system that enhances model performance by optimizing differentiable loss functions (Zhang *et al.*, 2023). These techniques are crucial in domains like cancer diagnosis, where feature selection can significantly impact classification accuracy and model interpretability (Al-Rajab, Lu and Xu, 2021).

Numerous studies have explored feature selection across different fields. For instance, Random Forest has been employed for feature selection in financial forecasting, demonstrating improved predictive performance with fewer features (Khanna, Kumar and Bhat, 2025). Similarly, XGBoost's built-in feature importance has been analyzed for its effectiveness in medical classification tasks, offering better interpretability and higher accuracy compared to traditional statistical approaches (Benghazouani, Nouh and Zakrani, 2024). These advancements highlight the versatility and effectiveness of embedded methods in handling complex datasets with numerous features.

Despite these advancements, inconsistencies persist regarding the effectiveness of various feature selection techniques. Some studies indicate that feature selection enhances accuracy, while others report minimal improvement or even performance degradation when critical features are removed (Attallah, 2025). The significance of feature selection in enhancing predictive accuracy has been highlighted by recent studies that show the successful application of machine learning models, such as Random Forest and XGBoost, in the categorization of lung cancer. Research has demonstrated that by using feature selection techniques to improve model performance, machine learning frameworks can accurately forecast clinical outcomes for patients with lung cancer, including hospital length of stay (LOS) and response to treatment (Alsinglawi *et al.*, 2022). Metaheuristic methods, including PSO and GA, have been explored to address these challenges, offering adaptive search capabilities that help optimize feature subsets dynamically (Shami *et al.*, 2022). In particular, GA has demonstrated effectiveness in handling imbalanced datasets, as it iteratively optimizes feature selection based on classification performance (Ileberi, Sun and Wang, 2022). Given that lung cancer

datasets often exhibit class imbalance, GA provides an efficient mechanism to optimize feature selection without compromising model accuracy, making it highly suitable for medical applications where dataset variability is a common challenge. This study further examines how Random Forest and XGBoost handle redundant and imbalanced medical features, highlighting their respective strengths RF's robustness to noisy attributes and XGBoost's sensitivity to structured patterns under controlled model configurations with GA based feature selection. Additionally, the computational efficiency of different selection methods varies, influencing their applicability in real-world medical scenarios where computational resources may be limited (Vijayalakshmi *et al.*, 2020). These variations highlight the importance of selecting appropriate feature selection techniques that balance accuracy, computational cost, and interpretability, ensuring robust classification performance for lung cancer detection and prognosis. (Sheth, Patil and Dhore, 2022).

This research aims to compare Random Forest and XGBoost for feature selection, evaluating their impact on model performance in lung cancer classification. Unlike previous studies that incorporate hyperparameter tuning, this research isolates the effect of feature selection by maintaining fixed model configurations. The dataset used is sourced from Kaggle, and the findings will provide insights into the most effective feature selection technique for classification tasks, potentially guiding future applications in similar high-stakes domains (Alsulami, 2024)(Bansal, Goyal and Choudhary, 2022).

2. RESEARCH METHOD

This study follows a structured approach, as outlined in Figure 1. Data collection, preprocessing, feature selection, model training, evaluation, and comparative analysis are all components of the study.

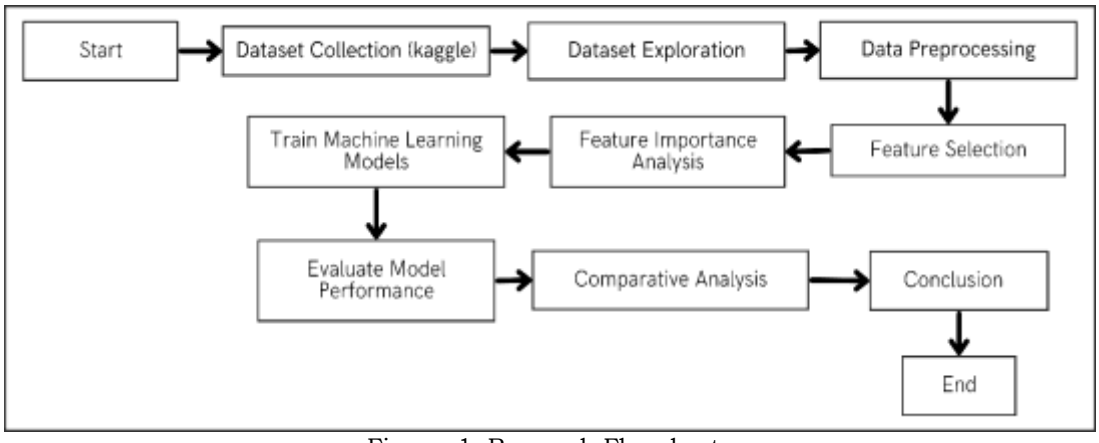

Figure. 1. Research Flowchart

2.1 Dataset Collection and Exploration

The dataset utilized in this work was obtained from Kaggle (Shantanu Garg, 2025) and included a number of attributes pertinent to tasks involving the categorization of lung cancer. Before preprocessing, dataset exploration is conducted to understand the data structure and identify potential issues, including class imbalance, missing values, and correlations (Safriandono, Setiadi, *et al.*, 2024). The key steps include: (a) Descriptive Statistics: Examining data distribution, mean, median, and standard deviation (Xu *et al.*, 2023); (b) Correlation Analysis: Identifying relationships between features to assess redundancy and feature significance (Zhang *et al.*, 2023); (c) Class Distribution Analysis:

Checking for class imbalances that may affect model performance (Safriandono, Setiadi, *et al.*, 2024)(Hammad *et al.*, 2024).

2.2 Data Preprocessing

Preprocessing steps ensure data consistency and enhance model performance (Razmjouei *et al.*, 2024)(Choudhry *et al.*, 2023). These include: (a) Handling Missing Values: Missing data is addressed using mean or mode imputation (Flyckt *et al.*, 2024). (b) Feature Normalization: Z-score normalization or Min-Max Scaling are used to standardize numerical features (Safriandono, Ignatius, *et al.*, 2024). © Coding Categorical Variables: Depending on the features, either label encoding or one-hot encoding is used to transform categorical data (Zhang *et al.*, 2020).

2.3 Feature Selection

Feature selection is performed to enhance model performance and reduce computational complexity (K.R.UTHAYAN1* and , B.NIVETHA2, 2021)(Budiman, 2025). Two methods are employed: (a) Random Forest Feature Selection: Evaluates feature importance using Gini impurity to select the most relevant features (Khanna, Kumar and Bhat, 2025). (b) XGBoost Feature Selection: Uses gradient boosting to assign importance scores to each feature based on its contribution to decision trees (Benghazouani, Nouh and Zakrani, 2024). To determine the optimal number of generations for the Genetic Algorithm (GA), this study explored several settings 10, 20, 30, 40, and 50 iterations to assess at which point convergence was achieved. Through this experimental setup, 50 generations were ultimately chosen, as performance improvements stabilized beyond the 40th generation while earlier configurations showed lower fitness scores or inconsistent convergence.

2.4 Feature Importance Analysis

Following the selection of the best features, additional research is done to evaluate each feature's effect on classification performance. To identify which features are most important for the prediction process, their importance is prioritized and displayed (Khanna, Kumar and Bhat, 2025).

2.5 Model Training & Evaluation

Both the Random Forest and XGBoost classifiers are trained using the chosen features. Both Random Forest and XGBoost classifiers were trained using the chosen features because of their robustness in managing complicated datasets and excellent feature selection capabilities. Random Forest was chosen for its stability across different feature selection methods and its resilience to feature variations, while XGBoost was selected for its superior predictive performance and ability to manage structured data efficiently (Alsinglawi *et al.*, 2022). The following measures are used to assess model performance:

Accuracy: Measures overall classification correctness (Hammad *et al.*, 2024)(Razmjouei *et al.*, 2024)(Benghazouani, Nouh and Zakrani, 2024).

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

Precision and Recall: Assesses how well the model detects positive cases while reducing false positives and false negatives. (Benghazouani, Nouh and Zakrani, 2024)

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: offers a trade off between recall and precision, which is especially helpful in datasets that are unbalanced (Benghazouani, Nouh and Zakrani, 2024).

$$F - 1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

For Random Forest, the prediction function is based on the majority vote of multiple decision trees:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), .., T_n(x))$$

Where $T_i(x)$ represent each of the ensemble's distinct decision trees.. For XGBoost, the model is formulated as:

$$\hat{y} = \sum_{K=1}^{K} f_k(x), f_k \in \mathcal{F}$$

where $\mathcal{F}$ is the space of regression trees, and each $f_k(x)$ represents a weak learner optimized using gradient boosting.

2.6 Comparative Analysis

The final step involves comparing the impact of feature selection using Random Forest and XGBoost. Performance metrics are analyzed to determine which method provides better classification accuracy and feature efficiency.

2.7 Conclusion

Using Random Forest and XGBoost, the study assesses how well feature selection enhances lung cancer categorization. The study emphasizes the significance of choosing pertinent features to improve forecast accuracy by contrasting model performance before and after feature selection. By guaranteeing that machine learning models maintain excellent classification performance while concentrating on the most informative qualities, the findings help improve medical diagnosis decision-making.

3. RESULTS AND DISCUSSIONS

3.1 Results

In medical classification tasks like lung cancer diagnosis, feature selection is particularly important for increasing model efficiency and interpretability. Reducing unnecessary and duplicated information allows the models to concentrate on the most instructive characteristics, which enhances predictive performance and generalization. This section compares the impact of Random Forest (RF) and XGBoost feature selection methods on classification performance. Additionally, we discuss the role of Genetic Algorithm (GA) in optimizing feature selection, particularly examining the effect of running 50 generations of GA on model outcomes.

a. Feature Selection Results

Prior research has demonstrated that combining multiple feature selection methods can yield better classification performance. Chen & Dhahbi (2021) showed that using a combination of Random Forest and XGBoost for feature selection in lung cancer classification could improve predictive accuracy, particularly in differentiating between adenocarcinoma and squamous cell carcinoma subtypes (Chen and Dhahbi, 2021). In this study, both RF and XGBoost selected overlapping features related to lung cancer risk factors, such as smoking, exposure to pollution, breathing issues, and family history

(Table 1). However, discrepancies were observed, where RF emphasized broader physiological indicators (e.g., age, energy level, stress-immune), while XGBoost focused on symptom-specific attributes (e.g., gender, finger discoloration, long-term illness, oxygen saturation). These findings align with the results of previous studies, suggesting that combining different feature selection techniques can enhance model robustness and interpretability (Chen and Dhahbi, 2021).

Table 1. Comparison of Selected Features by RF and XGBoost

No	Feature	RF	XGBoost
1	age	✓	X
2	gender	✓	X
3	smoking	✓	✓
4	finger_discoloration	✓	X
5	mental_stress	X	X
6	exposure_to_pollution	✓	✓
7	long_term_illness	X	X
8	energy_level	✓	✓
9	immune_weakness	✓	X
10	breathing_issue	✓	✓
11	alcohol_consumption	✓	X
12	throat_discomfort	✓	✓
13	oxygen_saturation	✓	X
14	chest_tightness	X	X
15	family_history	X	✓
16	smoking_family_history	✓	✓
17	stress_immune	✓	✓

As shown in Table 1, RF and XGBoost agreed on the importance of smoking, exposure to pollution, breathing issues, throat discomfort, and smoking family history, highlighting their relevance in lung cancer classification. However, RF selected additional features such as age, gender, finger discoloration, energy level, immune weakness, oxygen saturation, and stress-immune, suggesting that it captures a broader range of physiological and environmental factors. In contrast, XGBoost focused on a more compact subset, emphasizing energy level, family history, and stress-immune, which indicates a preference for symptom-specific indicators. The exclusion of chest tightness, oxygen saturation, and alcohol consumption from XGBoost's selection suggests that it prioritizes features with stronger direct predictive power, whereas RF leverages a wider set of attributes for improved generalization. This suggests that XGBoost prioritizes features with stronger direct predictive power, whereas RF leverages a wider set of variables for improved generalization. From a clinical perspective, this implies that RF may be more suitable for early risk assessment , where patient background and systemic health factors are crucial, while XGBoost may be better suited for diagnostic confirmation , where symptom-based features dominate clinical decision making.

b. Model Performance Evaluation

We compared Random Forest and XGBoost classifiers before and after GA-based feature selection in order to determine how feature selection affected model performance. The results are presented in Table 2.

Table 2. Model Performance Before and After GA-Based Feature Selection

No	Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
1	Random Forest Before GA	0.920	0.910	0.892	0.900	0.928
2	Random Forest After GA	0.921	0.910	0.894	0.902	0.924
3	XGBoost Before GA	0.912	0.894	0.889	0.892	0.927

4	XGBoost After GA	0.916	0.889	0.907	0.897	0.926

Following Genetic Algorithm (GA)-based feature selection, Random Forest fared better than XGBoost on the majority of evaluation criteria (Table 2). With the best accuracy (0.921), precision (0.910), and F1-score (0.902), RF demonstrated its exceptional capacity for accurate and balanced categorization. Although XGBoost demonstrated a slightly higher recall (0.907 vs. 0.894) and AUC-ROC (0.926 vs. 0.924), the overall improvements for RF were more pronounced. These results suggest that GA effectively enhanced RF's classification performance by refining feature selection and reducing noise. Meanwhile, XGBoost also benefited from GA-based feature selection but with relatively smaller gains, likely due to its reliance on gradient boosting rather than an inherent feature selection mechanism like RF's impurity-based evaluation.

c. Impact of Genetic Algorithm (GA) Iterations

In order to maximize the subset of chosen features while preserving classification accuracy, the Genetic Algorithm (GA) was integrated for feature selection. In this study, GA was executed for 50 generations, balancing computational efficiency with the need for optimization. The choice of 50 generations was based on preliminary experimentation, which indicated that increasing the number of generations beyond this point resulted in marginal improvements in feature selection quality and classification performance.

Figure 2 illustrates the convergence of GA over 50 generations for two machine learning models: Random Forest (RF) and XGBoost (XGB). The blue line represents the best fitness score progression for RF, while the red line represents XGB. The dotted vertical line at the 40th generation indicates the stabilization point, where the fitness scores no longer show significant improvements.

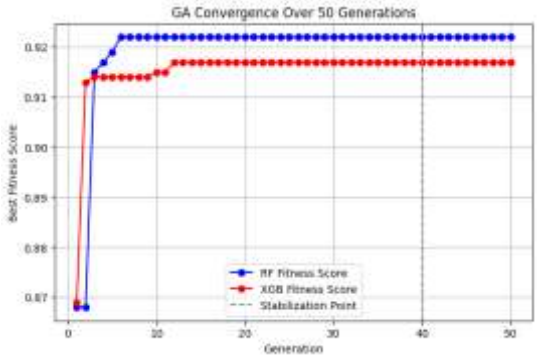

Figure 2: Graph showing convergence of selected feature subsets over 50 generations

From the graph, we can observe that: (a) XGBoost (red) experiences rapid improvement in the first 10 generations before gradually stabilizing around generation 20 with a fitness score of approximately 0.915. (b) Random Forest (blue) exhibits a slightly steeper initial increase and reaches a marginally higher fitness score (~0.923) around generation 10, maintaining stability throughout the remaining iterations. (c) The stabilization point at generation 30 suggests that additional generations beyond this do not contribute to meaningful performance gains.

By running GA for 50 generations, we ensured that the selected feature subset had stabilized after approximately 40 generations, indicating that further increasing the number of iterations would not yield significant improvements while unnecessarily increasing computational costs. The number of generations in the Genetic Algorithm was set to 50 based on exploratory convergence analysis. As illustrated in Figure 2, fitness scores began to stabilize around the 40th generation, suggesting that extending the iterations further would not yield substantial improvements in feature quality. This choice balances the risk of overfitting with computational efficiency, ensuring that GA achieves optimal feature selection without incurring unnecessary processing costs. The

optimized feature sets obtained through GA were then evaluated using RF and XGBoost, demonstrating improvements in precision, recall, and AUC-ROC scores (as illustrated in Figure 3).

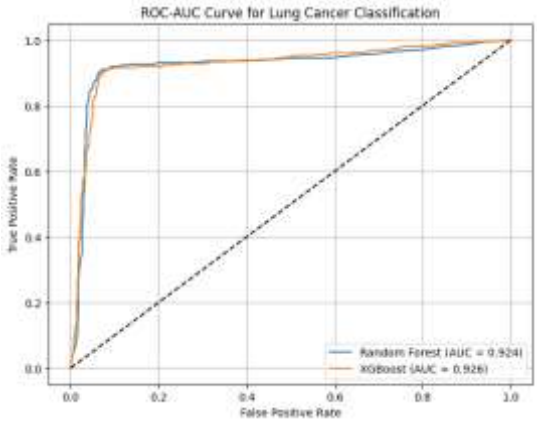

Figure 3: ROC-AUC Curve for Lung Cancer Classification using RF and XGBoost

The Random Forest (blue) and XGBoost (orange) models' performance in classifying lung cancer is shown by the ROC-AUC curve above. This graph illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across various classification thresholds. According to the Area Under the Curve (AUC) values, XGBoost (AUC = 0.926) slightly outperforms Random Forest (AUC = 0.924). Both models exhibit similar performance, reflecting their strong ability to distinguish between positive and negative cases. The high AUC values, exceeding 0.92, suggest that these models are highly effective in predicting lung cancer cases with minimal misclassification. In real-world applications, such models are crucial for early cancer detection, where minimizing false negatives is particularly important. The findings indicate that while XGBoost provides a marginal improvement in classification accuracy, both models are highly reliable for medical diagnosis tasks. These results demonstrate that, even though GA is an effective technique for feature selection optimization, too many iterations may result in declining returns. To balance accuracy, interpretability, and computational efficiency for real-world applications, the right number of generations must be chosen.

3.2 Discussion

A The results indicate that Random Forest is a more suitable choice for feature selection in lung cancer classification compared to XGBoost. Several factors contribute to this conclusion:

a. Robustness in Handling Noisy and Correlated Features

Because it ranks feature importance based on impurity reduction, RF is well known for its ability to manage datasets containing redundant and irrelevant features. Prior research has demonstrated that tree-based feature selection methods, such as RF, effectively reduce the impact of noisy and correlated features, leading to more stable and interpretable models in cancer classification (Mohamed Ebrahim, 2023). XGBoost, while effective in structured datasets, is more sensitive to noisy features, which can affect feature ranking. Although gradient boosting algorithms such as XGBoost excel in structured tabular data, they may overfit to noise when feature selection is not carefully optimized (Mohamed Ebrahim, 2023).

b. Better Generalization and Stability

The selected features by RF align more closely with known risk factors for lung cancer, such as smoking, exposure to pollution, and breathing issues, indicating its ability to generalize well. XGBoost, on the other hand, introduced features like gender and finger discoloration, which may not be as strongly linked to lung cancer classification. This aligns with previous studies suggesting that tree boosting models may assign importance to weakly correlated features if their contribution to model variance is high (Mohamed Ebrahim, 2023).

c. Improved Model Performance with GA

RF demonstrated more significant improvements after GA-based feature selection, highlighting its synergy with evolutionary optimization techniques.

The improvement in RF's AUC-ROC and recall indicates a stronger ability to differentiate between lung cancer and non-lung cancer cases, a critical factor in medical diagnosis. These findings further validate that combining GA with ensemble feature selection methods can lead to improved classification accuracy and model interpretability (Mohamed Ebrahim, 2023) development of a prototype enterprise blockchain system using Hyperledger Fabric.

A two-tiered architecture for Central Bank Digital Currency (CBDC) is proposed within the existing financial infrastructure, designed specifically for a permissioned blockchain network-based CBDC (Khanna, Kumar and Bhat, 2025). The use of two layers is also in line with the conceptual design designed by BI (Kaur and Kumari, 2022) (Nadimi-shahraki, Zamani and Mirjalili, 2022). The initial layer, referred to as the distributional layer, encompasses wholesale CBDC and defines various smart contracts within the Permissioned Blockchain Network (PBN). Commercial banks, mirroring the central bank, uphold a blockchain node in the PBN to establish legitimacy. The Certificate Authority (CA) assumes a vital role in distributing digital certificates to all participating banks in the PBN, validating transactions related to tokens, and ensuring secure communication. Significantly, the central bank maintains no direct interaction with end-users but instead authorizes commercial banks to allocate token-based accounts to end-users.

This proposed architecture offers several advantages over existing systems like Real-Time Gross Settlement (RTGS) and Immediate Payment Service (IMPS) commonly employed by central banks. Notably, it eliminates a single point of failure by introducing the PBN for wholesale CBDC, enabling direct interaction between commercial banks without reliance on existing interbank settlement systems (e.g., RTGS), thereby achieving zero downtime. Scalability and cost considerations are addressed through the utilization of smart contract logic and a permissioned blockchain environment, aiming to reduce overall system costs and facilitate increased transaction volume as the increasing of commercial banks. The use of blockchain technology in the distribution layer is emphasized for its capacity to introduce transparency and trust of all banks by securing a distributed immutable ledger. Looking forward, potential enhancements for the proposed CBDC system architecture include implementing hybrid encryption cryptography for the PBN to safeguard against potential quantum computing attacks.

Regulation consent regarding consumer data protection is very high because Rupiah digital uses blockchain technology so it is necessary to ensure that transactions in it can only be known by the authorities. In discussions related to regulations, the views of all participants felt that the P2SK Law was not enough to give privacy because the validating node could know the costumer. BI still conducting a research to obtain the best ways of transaction which can’t be seen by banks in order to gain privacy for end-user.

4. CONCLUSION

Provide a statement that what is expected, as stated in the "Introduction" chapter can Feature selection enhances machine learning models for lung cancer diagnosis by reducing computational costs, improving interpretability, and boosting predictive performance. In this study, the implementation of Genetic Algorithm (GA) significantly improved model accuracy compared to using all features, as evidenced by increased AUC-ROC scores, precision, and recall for both Random Forest and XGBoost. Before GA, the models exhibited lower performance due to the presence of redundant or less relevant features. After optimization, the refined feature subset led to a more efficient and accurate classification process. This emphasizes how crucial organized feature selection is for problems involving medical classification. Future research could explore hybrid feature selection techniques combining filter, wrapper, and embedded methods, as well as integrating deep learning-based approaches like autoencoders or attention mechanisms to enhance feature interpretability and robustness. Additionally, optimizing hyperparameters alongside feature selection and applying this approach to multi-modal datasets, including radiology images and genomic data, could further validate its effectiveness in clinical applications.

The integration of GA in feature selection shows potential for real-time diagnosis systems, particularly in balancing accuracy and computational efficiency. While GA improves model performance by identifying optimal feature subsets, its iterative nature may pose challenges in resource-constrained healthcare settings. However, the reduced feature set achieved through GA can lead to faster inference times and lower hardware requirements, making it feasible for deployment in clinical environments with limited computing power. Future work should explore lightweight GA implementations or hybrid approaches that combine filter/embedded methods to further improve efficiency without compromising predictive performance.

REFERENCES

Al-Rajab, M., Lu, J. and Xu, Q. (2021) 'A framework model using multifilter feature selection to enhance colon cancer classification', *PLoS ONE*, 16(4 April). Available at: https://doi.org/10.1371/journal.pone.0249094.

Alsinglawi, B. *et al.* (2022) 'An explainable machine learning framework for lung cancer hospital length of stay prediction', *Scientific Reports*, 12(1), pp. 1–10. Available at: https://doi.org/10.1038/s41598-021-04608-7.

Alsulami, A.A. (2024) 'An Efficient Model for Lung Cancer Detection through the Integration of Genetic Algorithm and Machine Learning', *Engineering, Technology & Applied Science Research*, 14(6), pp. 18792–18798.

Attallah, O. (2025) 'Lung and Colon Cancer Classification Using Multiscale Deep Features Integration of Compact Convolutional Neural Networks and Feature Selection', *Technologies*, 13(2), pp. 1–28. Available at: https://doi.org/10.3390/technologies13020054.

Bansal, M., Goyal, A. and Choudhary, A. (2022) 'A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning', *Decision Analytics Journal*, 3(May), p. 100071. Available at: https://doi.org/10.1016/j.dajour.2022.100071.

Benghazouani, S., Nouh, S. and Zakrani, A. (2024) 'Enhancing breast cancer diagnosis: a comparative analysis of feature selection techniques', *IAES International Journal of Artificial Intelligence*, 13(4), pp. 4312–4322. Available at: https://doi.org/10.11591/ijai.v13.i4.pp4312-4322.

Budiman, E. (2025) *Lung Cancer Git*.

Chen, J.W. and Dhahbi, J. (2021) 'Lung adenocarcinoma and lung squamous cell carcinoma cancer classification , biomarker identification , and gene expression analysis using overlapping feature selection methods', *Scientific Reports*, pp. 1–15. Available at: https://doi.org/10.1038/s41598-021-92725-8.

Choudhry, I.A. *et al.* (2023) 'Hybrid Diagnostic Model for Improved COVID-19 Detection in Lung Radiographs Using Deep and Traditional Features', *Biomimetics*, 8(5), pp. 1–19. Available at: https://doi.org/10.3390/biomimetics8050406.

Flyckt, R.N.H. *et al.* (2024) 'Pulmonologists-Level lung cancer detection based on standard blood test results and smoking status using an explainable machine learning approach', *Scientific Reports*, pp. 1–11. Available at: https://doi.org/10.1038/s41598-024-82093-4.

Göltepe, Y. (2021) 'Performance of lung cancer prediction methods using different classification algorithms', *Computers, Materials and Continua*, 67(2), pp. 2015–2028. Available at: https://doi.org/10.32604/cmc.2021.014631.

Hammad, M. *et al.* (2024) 'Automated lung cancer detection using novel genetic TPOT feature optimization with deep learning techniques', *Results in Engineering*, 24(October), p. 103448. Available at: https://doi.org/10.1016/j.rineng.2024.103448.

Ileberi, E., Sun, Y. and Wang, Z. (2022) 'A machine learning based credit card fraud detection using the GA algorithm for feature selection', *Journal of Big Data* [Preprint]. Available at: https://doi.org/10.1186/s40537-022-00573-8.

K.R.UTHAYAN1*, S.M. and , B.NIVETHA2, S.D. (2021) 'OPTIMISED FEATURE SELECTION FOR EARLY CANCER DETECTION', *Original scientific article*, 53, pp. 985-996,.

Kaur, H. and Kumari, V. (2022) 'Predictive modelling and analytics for diabetes using a machine learning approach', *Applied Computing and Informatics*, 18(1–2), pp. 90–100. Available at: https://doi.org/10.1016/j.aci.2018.12.004.

Khanna, D., Kumar, A. and Bhat, S.A. (2025) 'Volatile Organic Compound for the Prediction of Lung Cancer by using Ensembled Machine Model and Feature Selection', *IEEE Access*, PP(January), p. 1. Available at: https://doi.org/10.1109/ACCESS.2025.3527027.

Mohamed Ebrahim, A.A.H.S. and S.M. (2023) 'Accuracy Assessment of Machine Learning Algorithms Used to Predict Breast Cancer', *University of Cambridge* [Preprint].

Nadimi-shahraki, M.H., Zamani, H. and Mirjalili, S. (2022) 'Enhanced whale optimization algorithm for medical feature selection : A COVID-19 case study', *Computers in Biology and Medicine*, 148(June), p. 105858. Available at: https://doi.org/10.1016/j.compbiomed.2022.105858.

Razmjouei, P. *et al.* (2024) 'Metaheuristic-Driven Two-Stage Ensemble Deep Learning for Lung/Colon Cancer Classification', *Computers, Materials and Continua*, 80(3), pp. 3855–3880. Available at: https://doi.org/10.32604/cmc.2024.054460.

Safriandono, A.N., Setiadi, D.R.I.M., *et al.* (2024) 'Analyzing Quantum Feature Engineering and Balancing Strategies Effect on Liver Disease Classification', *Journal of Future Artificial Intelligence and Technologies*, 1(1), pp. 51–63. Available at: https://doi.org/10.62411/faith.2024-12.

Safriandono, A.N., Ignatius, D.R., *et al.* (2024) 'Journal of Future Artificial Intelligence Analyzing Quantum Feature Engineering and Balancing Strategies Effect on Liver Disease Classification', *Future Techno Science* [Preprint].

Shami, T.M. *et al.* (2022) 'Particle Swarm Optimization: A Comprehensive Survey', *IEEE Access*, 10, pp. 10031–10061. Available at: https://doi.org/10.1109/ACCESS.2022.3142859.

Shantanu Garg (2025) *Lung Cancer Prediction Dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/shantanugarg274/lung-cancer-prediction-dataset (Accessed: 17 March 2025).

Sheth, P.D., Patil, S.T. and Dhore, M.L. (2022) 'Evolutionary computing for clinical dataset classification using a novel feature selection algorithm', *Journal of King Saud University - Computer and Information Sciences*, 34(8), pp. 5075–5082. Available at: https://doi.org/10.1016/j.jksuci.2020.12.012.

Too, J. and Mirjalili, S. (2021) 'Knowledge-Based Systems A Hyper Learning Binary Dragonfly Algorithm for Feature Selection : A COVID-19 Case Study', *Knowledge-Based Systems*, 212, p. 106553. Available at: https://doi.org/10.1016/j.knosys.2020.106553.

Vijayalakshmi, S. *et al.* (2020) 'Multi-modal prediction of breast cancer using particle swarm optimization with non-dominating sorting', *International Journal of Distributed Sensor Networks*, 16(11). Available at: https://doi.org/10.1177/1550147720971505.

Xu, K. *et al.* (2023) 'AI Body Composition in Lung Cancer Screening: Added Value Beyond Lung Cancer Detection', *Radiology*, 308(1). Available at: https://doi.org/10.1148/radiol.222937.

Zhang, J.U.N. *et al.* (2020) 'Cyber Resilience in Healthcare Digital Twin on Lung Cancer', *IEEE Access*, 8. Available at: https://doi.org/10.1109/ACCESS.2020.3034324.

Zhang, Y.P. *et al.* (2023) 'Artificial intelligence-driven radiomics study in cancer: the role of feature engineering and modeling', *Military Medical Research*, 10(1), pp. 1–33. Available at: https://doi.org/10.1186/s40779-023-00458-8.