OPEN ACCESS
International Journal of Advances in Artificial Intelligence and Machine Learning Vol 2 No 1 (2025) https://doi.org/10.58723/ijaaiml.v2i1.378
Check for updates
Research Article

A Comparative Evaluation of Predictive Models for Lung Cancer: Insights from Logistic Regression, Naive Bayes, and Random Forest

Received: January 19, 2025	Revised: March 08, 2025	Accepted: March 14, 2025	Publish: March 15, 2025

Muhammad Hafiz Kurniawan*, Misinem

Abstract:
This study aims to evaluate the performance of three machine learning models-Logistic Regression, Naive Bayes, and Random Forest-in predicting lung cancer using a publicly available dataset from Kaggle. The data used included demographic information, risk factors, and diagnostic imaging features, with significant class imbalance between benign and malignant cases. To address this imbalance, the Synthetic Minority Sampling Technique (SMOTE) was applied. In addition, Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) were used for dimensionality reduction and feature selection to improve model performance. The results showed that Random Forest, especially when combined with PCA, outperformed the other models with the highest accuracy of 96.77% and a balanced F1 score of 0.50 for the minority class. Although Logistic Regression achieved high accuracy, it was less effective in predicting minority classes, especially when combined with RFE. Meanwhile, Naive Bayes showed moderate performance but was limited by the assumption of feature independence. The application of SMOTE significantly improved the model's ability to handle class imbalance, while PCA proved more effective than RFE in improving model performance. This study highlights the importance of selecting appropriate machine learning models and preprocessing techniques for lung cancer prediction. Random Forest, with its ability to model complex relationships and handle imbalanced data, emerged as the most effective model for this task. These findings underscore the potential of machine learning in medical diagnostics and provide valuable insights for future research.

Keywords: Logistic Regression, Lung Cancer, Machine Learning, Naive Bayes, Random Forest.

1. INTRODUCTION

Lung cancer remains one of the biggest medical challenges, with a higher mortality rate than any other type of cancer (Li et al., 2025). The disease is often detected at an advanced stage, worsening patient prognosis and limiting available treatment options. In an effort to improve survival rates, early detection is crucial. Unfortunately, conventional methods of diagnosis, such as medical imaging and tissue biopsy, have their limitations. Besides being expensive and time-consuming, these procedures can also be invasive and carry risks for patients. Therefore, a more efficient and accurate approach is needed to improve the detection of lung cancer at an early stage (Parisi et al., 2024). Advances in artificial intelligence (AI) and machine learning (ML) have opened up new opportunities in the field of medical diagnostics. ML models are capable of analyzing large amounts of data with complex patterns, often surpassing both the accuracy of traditional statistical methods and the level of human expertise in certain diagnostic tasks (Hemmer et al., 2023). The main advantage of ML lies in its ability to identify subtle patterns in medical data that may be difficult for doctors to recognize manually. However, with many ML models available, a key challenge is determining which algorithm is most effective in detecting lung cancer, given the heterogeneity of clinical data and variation in disease presentation (Richens et al., 2020).

To address this challenge, this study will explore and compare three commonly used ML models in medical classification: Logistic Regression, Naive Bayes, and Random Forest. Logistic Regression is known for its high interpretability, which allows medical personnel to understand the main risk factors that contribute to a diagnosis (Linardatos & Papastefanopoulos, 2021). On the other hand, Naive Bayes offers an efficient probabilistic approach in processing small datasets assuming independence between features (Chen et al., 2020). Meanwhile, Random Forest stands out in handling complex relationships between variables with a high degree of accuracy, making it a strong

choice for the analysis of more complex medical data (Simon et al., 2023).

This study utilized a dataset obtained from Kaggle, which included demographic information, risk factors, as well as preliminary imaging results from individuals who had undergone lung cancer screening. One of the main challenges found in this dataset is the imbalance of the class distribution, which risks causing bias in the model predictions. When ML models are trained with an imbalanced dataset, they often become more inclined to classify cases into the majority category, thus reducing sensitivity to rarer cancer cases. Therefore, data balancing techniques will be applied to ensure that the model can effectively learn from both classes and produce more accurate predictions (Thabtah et al., 2019).

In addition to selecting an appropriate model, this research will also use feature selection techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) to improve prediction efficiency (Hermiati et al., 2024). By reducing the dimensionality of the data and highlighting the most relevant features, it is expected that the model can work more optimally without losing important information related to lung cancer diagnosis.

As a final step, the entire process of model training and evaluation will be rigorously documented to ensure reproducibility of results and fair comparison between models. Thus, this research not only contributes to selecting the most accurate ML model for detecting lung cancer, but also provides deeper insights into how artificial intelligence can change the landscape of medical diagnosis in the future. If applied correctly, this technology could potentially save more lives by providing faster, cheaper and more accurate early detection than current conventional methods.

2. MATERIAL AND METHOD

To Dataset Acquisition and Preprocessing

The lung cancer dataset utilised for this study was sourced from Kaggle. It includes comprehensive patient information such as demographics, smoking history, family cancer history, and initial diagnostic test results from CT scans and MRIs. Each record is labelled as "benign" or "malignant," providing a clear target for supervised learning.

Data Preprocessing Steps

1. Handling Missing Values: Missing values were addressed using imputation techniques. Numerical features were imputed with the mean or median values, while categorical features were imputed using the mode.
2. Addressing Class Imbalance: The dataset exhibited a significant imbalance, with a predominance of malignant samples. The Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data to balance the dataset, thereby enhancing the model's learning from the minority class.
3. Feature Standardization: Numerical features such as age were standardised using the StandardScaler to normalise the data, ensuring that all features contributed equally to model training.

Feature Selection Techniques

To improve model performance and computational efficiency, two feature selection techniques were employed:

1. Principal Component Analysis (PCA): PCA was used to reduce the dimensionality of the dataset while retaining the most informative features. This involved transforming the original features into a set of orthogonal components that captured the maximum variance in the data (Uddin et al., 2020).
2. Recursive Feature Elimination (RFE): RFE was utilised to prune less significant features selectively. This technique works by recursively training the model and removing the least important features, based on their weights, until the desired number of features is retained (Ma et al., 2024).

Model Implementation

Three machine-learning models were selected based on their applicability to medical diagnostics:

1. Logistic Regression: A linear model providing probabilistic outputs, favoured for its interpretability in medical settings (Boateng & Abaye, 2019).
2. Naive Bayes: A probabilistic model that assumes independence among predictors, known for its computational efficiency and suitability for small datasets (Nakhipova et al., 2024).
3. Random Forest: An ensemble method that uses multiple decision trees to improve predictive accuracy and robustness, effectively handling complex interactions and non-linear relationships (Ahmad et al., 2024).

Each model was trained using both the complete feature set and the subsets selected by PCA and RFE.

Model Evaluation

Model performance was assessed using several key metrics:

1. Accuracy: The ratio of correctly predicted instances to the total dataset.
2. Precision, Recall, and F1-Score: Calculated for each class to evaluate the models' ability to identify each class accurately.
3. Confusion Matrix: Used to visualise the performance of the models in terms of true positives, true negatives, false positives, and false negatives.

Experimental Setup

1. Data Splitting: The dataset was divided into training (80%) and testing (20%) sets to validate the models on unseen data.
2. Reproducibility: A fixed random seed was used during all phases of the experiment to ensure that results were reproducible.

Comparative Analysis

Each model's performance was evaluated under uniform conditions using both PCA and RFE feature selection techniques. This comparative analysis aimed to determine the most effective model and feature selection combination for predicting lung cancer based on the dataset characteristics.

The study provided a comprehensive examination of the effectiveness of various preprocessing techniques, feature selection methods, and machine learning models, offering insights into their applicability for lung cancer diagnostics in a clinical setting (Mani & Rajaguru, 2024).

3. RESULT AND DISCUSSION

To The study follows a systematic approach to data preprocessing, model implementation, and evaluation, ensuring robust and reproducible results. The dataset used in this research is sourced from Kaggle. It includes patient demographics, smoking history, family cancer history, and results from initial diagnostic tests such as CT scans and MRIs. Each record in the dataset is labelled as either "benign" or "malignant," providing a clear target for supervised learning models. The dataset was obtained from Kaggle with the URL https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer.

Missing values in the dataset are handled using appropriate imputation techniques. For numerical features, the mean or median is used, while categorical features are imputed using the mode. The dataset exhibits a significant class imbalance, with a majority of samples belonging to the malignant class. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training data. SMOTE generates synthetic samples for the minority class, ensuring balanced learning and improving the model's ability to predict minority class instances. Additionally, numerical features such as age are standardised using StandardScaler to ensure that all features contribute equally to the model training process.

Two feature selection techniques are employed to enhance model performance and reduce computational complexity: Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). PCA is applied to reduce the dimensionality of the dataset while retaining the most informative features. This technique transforms the original features into a set of orthogonal components, capturing the maximum variance in the data. On the other hand, RFE is used to iteratively eliminate less significant features, retaining the most relevant subset. This technique works by recursively training the model and removing the least essential features until the desired number of features is achieved. Both PCA and RFE are applied to the dataset to evaluate their impact on model performance.

Three widely used machine learning models are implemented and evaluated: Logistic Regression, Naive Bayes, and Random Forest. Logistic Regression is a linear model that provides probabilistic predictions and is highly interpretable, making it particularly useful in medical settings where understanding the influence of different variables is crucial. Naive Bayes is a probabilistic model based on Bayes' theorem, assuming independence among predictors. It is computationally efficient and performs well with small datasets. Random Forest, an ensemble method, constructs multiple decision trees and aggregates their results to enhance predictive accuracy and robustness. Known for its ability to model complex interactions and nonlinear relationships, Random Forest is a strong candidate for handling the complexities of medical data. Each model is trained and tested using both PCA and RFE to evaluate the impact of feature selection techniques on performance.

The performance of the models is evaluated using several metrics to ensure a comprehensive assessment. Accuracy measures the overall correctness of predictions, calculated as the ratio of correctly classified instances to the total number of

instances. The classification report provides a detailed evaluation of model performance, including precision, recall, and F1-score for each class. Precision represents the ratio of true positives to the total number of predicted positives, while recall measures the ratio of true positives to the total number of actual positives. The F1-score, the harmonic mean of precision and recall, provides a balanced measure of model performance. Additionally, the confusion matrix is used to visualise the number of true positives, true negatives, false positives, and false negatives, offering further insights into model performance.

The dataset is split into training and testing sets using an 80:20 ratio, ensuring that the models are evaluated on unseen data. Each model is trained on the training set and evaluated on the testing set. To ensure reproducibility, the random seed is fixed throughout the experiment. The performance of Logistic Regression, Naive Bayes, and Random Forest is compared under uniform conditions, with and without the application of PCA and RFE. This comparative analysis aims to identify the most effective model for lung cancer prediction and to highlight the strengths and weaknesses of each approach in the context of lung cancer diagnostics.

The study evaluated the performance of three machine learning models—Logistic Regression, Naive Bayes, and Random Forest—using two feature selection techniques, Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). The models were trained and tested on a lung cancer dataset obtained from Kaggle, which was pre-processed to handle missing values, address class imbalance using SMOTE, and standardise numerical features, as shown in Table 1.

Table 1. Comparison Result of each Classification Model

Model	Type	Class 0 F1	Class 1 F1	Accuracy	Macro Avg	Weight Avg
Logistic Regression	RFE	0.00	**0.98**	**0.97**	0.49	0.95
	PCA	0.40	0.97	0.95	0.69	**0.96**
Naïve Bayes	RFE	0.40	0.97	0.95	0.69	**0.96**
	PCA	0.33	0.97	0.94	0.65	0.95
Random Forest	RFE	0.33	0.97	0.94	0.65	0.95
	PCA	**0.50**	**0.98**	**0.97**	**0.74**	0.97

Logistic Regression with PCA achieved an accuracy of 95.16%, but its ability to classify the minority class (benign) was limited, with an F1-score of 0.40. With RFE, accuracy improved to 96.77%, but the model ultimately failed to detect benign cases, highlighting its limitations in handling imbalanced data, as shown in Figure 1.

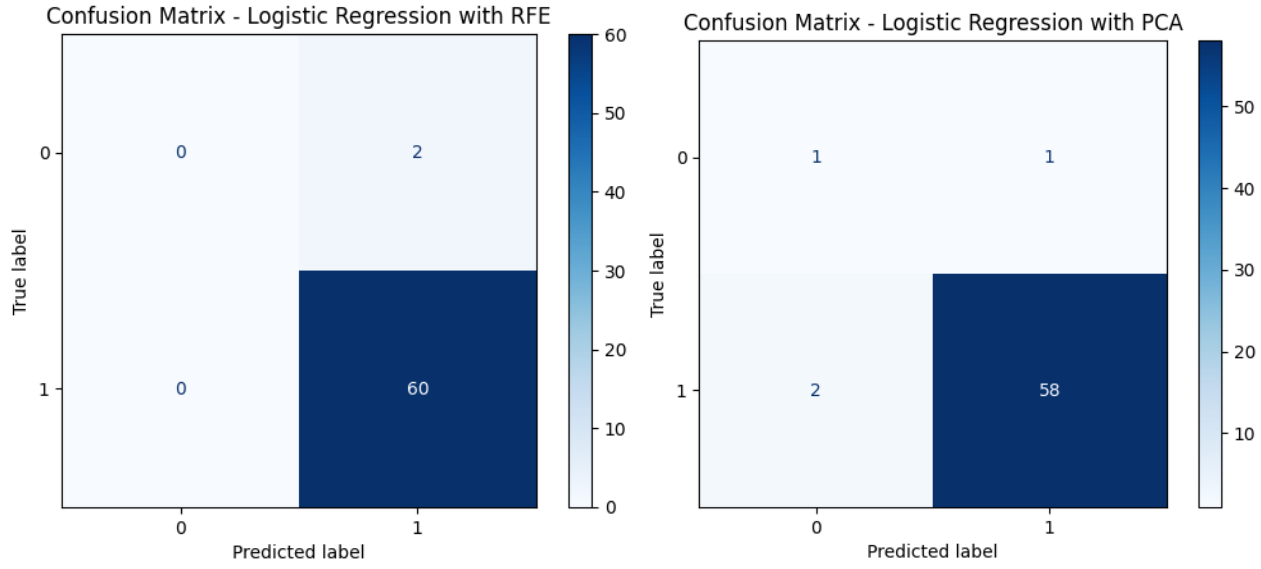

Figure 1. Logistic Regression confusion matrix results

Naive Bayes with PCA had an accuracy of 93.55% with an F1-score of 0.33 for the minority class, indicating weak performance in detecting benign cases. With RFE, accuracy increased to 95.16%, and

the F1-score for the minority class improved to 0.40, showing a slight enhancement in handling class imbalance, as shown in Figure 2.

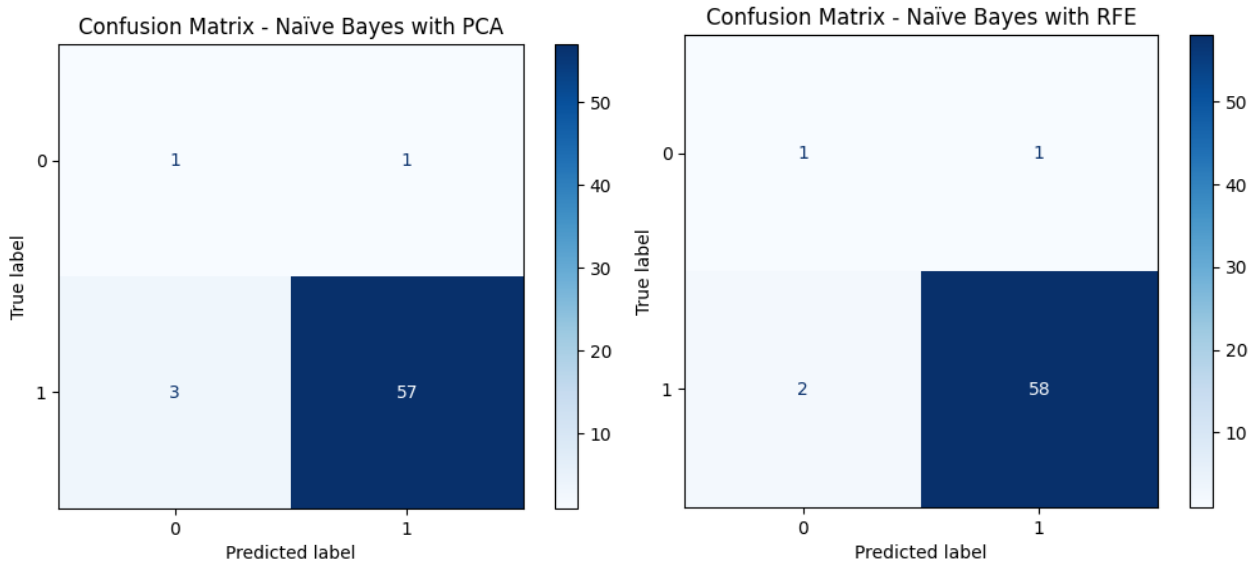

Figure 2. Naïve Bayes confusion matrix results

Random Forest with PCA achieved the highest accuracy of 96.77% and demonstrated the best balance in detecting both classes, with an F1-score of 0.50 for the minority class. However, with RFE, accuracy dropped to 93.55%, indicating that PCA was more effective than RFE in enhancing Random Forest's performance, as shown in Figure 3.

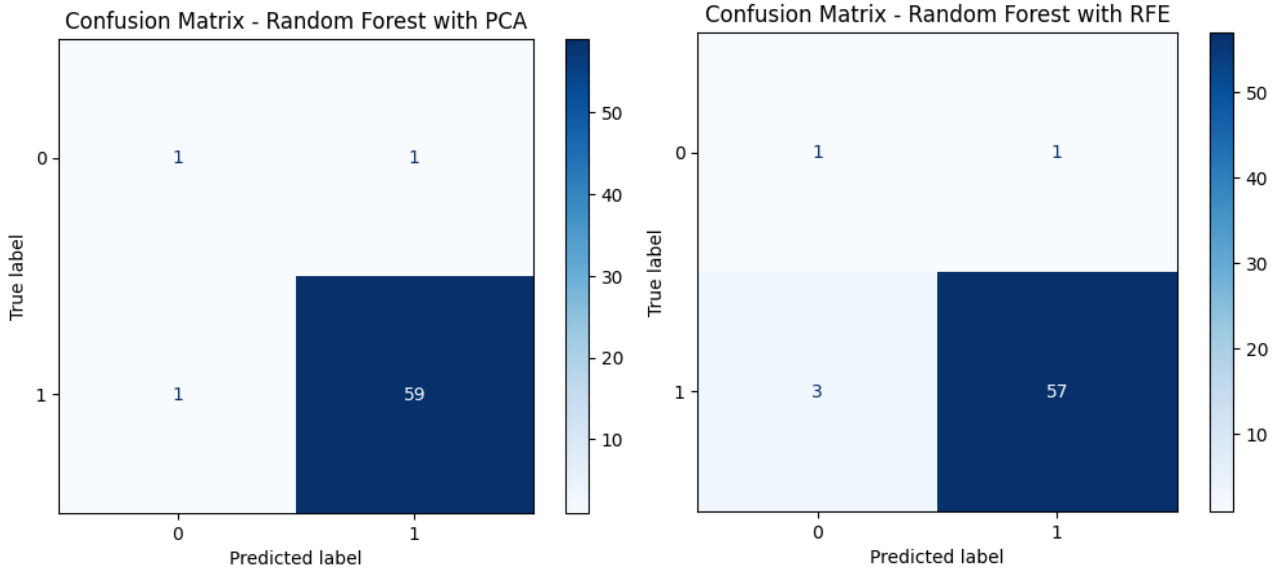

Figure 3. Random Forest confusion matrix results

Based on these findings, Random Forest with PCA is recommended as the best model, as it achieves the optimal balance between accuracy and minority class detection.

The results demonstrate that Random Forest outperformed Logistic Regression and Naive Bayes in terms of overall accuracy and minority class prediction. When combined with PCA, Random Forest achieved the highest accuracy (96.77%) and the most balanced F1-score (0.50) for the minority class. This superior performance can be attributed to Random Forest's ability to model complex, non-linear relationships and its robustness to overfitting. The application of PCA further enhanced its performance

by reducing dimensionality and retaining the most informative features.

In contrast, Logistic Regression struggled to predict the minority class effectively, particularly when combined with RFE. Despite achieving high accuracy (96.77%), the model failed to correctly classify any instances of the minority class, resulting in a precision and recall of 0.00. This highlights the limitations of Logistic Regression in handling imbalanced datasets, even with advanced feature selection techniques.

Naive Bayes showed moderate performance, with slight improvements when RFE was applied. However, its assumption of feature independence likely limited its effectiveness, as real-world medical data often contains correlated features. The model's precision and recall for the minority class remained low, indicating that Naive Bayes may not be the most suitable choice for this type of dataset.

The application of SMOTE played a crucial role in addressing the class imbalance, enabling the models to learn from both classes more effectively. Without SMOTE, the models would have been heavily biased toward the majority class, resulting in poor performance for the minority class. Additionally, the use of PCA and RFE demonstrated that feature selection techniques can significantly impact model performance. PCA, in particular, proved to be more effective than RFE for Random Forest, as it retained the most informative features while reducing dimensionality.

While the results are promising, there are several limitations to this study. First, the dataset used is relatively small, which may limit the generalizability of the findings. Second, the class imbalance, even after applying SMOTE, remains a challenge, as the minority class is still underrepresented. Third, the study focused on three models and two feature selection techniques, and future work could explore additional models (e.g., XGBoost, Support Vector Machines) and advanced techniques for handling imbalanced data.

This study highlights the importance of selecting appropriate machine learning models and preprocessing techniques for lung cancer prediction. Random Forest, particularly when combined with PCA, demonstrated superior performance in terms of accuracy and minority class prediction. Logistic Regression and Naive Bayes, while computationally efficient, were less effective in handling the dataset's imbalance and complexity. These findings underscore the potential of machine learning in medical diagnostics and provide valuable insights for future research in this domain.

4. CONCLUSION

This This study systematically evaluated the performance of three machine learning models Logistic Regression, Naive Bayes, and Random Forest on a Kaggle-sourced lung cancer dataset, which was rigorously pre-processed to handle missing values and class imbalance and standardised to ensure uniform contribution across features. The application of feature selection techniques, specifically Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), provided a focused analysis of their impact on the model's predictive accuracy and computational efficiency.

The findings reveal that Random Forest, combined with PCA, emerged as the most effective model, demonstrating superior accuracy (96.77%) and a balanced F1-score (0.50) for the minority class. This model's strength lies in its ability to manage complex, non-linear relationships within the data, which is enhanced further by PCA's dimensionality reduction, enabling the retention of the most informative features.

Conversely, Logistic Regression and Naive Bayes exhibited limitations in handling the dataset's inherent class imbalance and complexity. Logistic Regression, although achieving high accuracy, failed to effectively classify the minority class when combined with RFE, underscoring its challenges with imbalanced data. Naive Bayes, while slightly improved with the application of RFE, still struggled due to its assumption of feature independence, which is often violated in medical datasets.

The use of SMOTE proved critical in balancing the classes, enabling more effective learning from the minority class and improving overall model performance. This study underscores the importance of using appropriate preprocessing techniques and feature selection methods to enhance model training and prediction accuracy.

Furthermore, while the results are promising, they also highlight the constraints of using a limited dataset size and the persistent issue of class imbalance even after employing SMOTE. Future research could expand upon this work by exploring additional models such as XGBoost or Support Vector Machines and employing more advanced techniques for handling imbalanced data to improve model robustness and generalizability.

In conclusion, the integration of machine learning into medical diagnostics, particularly in lung cancer prediction, offers significant potential. This study

contributes valuable insights into the selection and application of machine learning models and preprocessing techniques, which can aid in the development of more accurate and reliable diagnostic tools in the healthcare industry.

AUTHOR INFORMATION

Corresponding Authors

Muhammad Hafiz Kurniawan, Universitas Sriwijaya, Palembang, Indonesia.
https://orcid.org/0009-0007-5069-0811
Email: kurni.hafiz2002@gmail.com

Authors

Misinem, Universitas Bina Darma, Palembang, Indonesia.
https://orcid.org/0000-0002-7946-4582
Email: misinem@binadarma.ac.id

REFFERENCE

Ahmad, A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems : Combination , implementation and evaluation. *Expert Systems With Applications*, *244*(May 2023), 122778. https://doi.org/10.1016/j.eswa.2023.122778

Boateng, E. Y., & Abaye, D. A. (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research. *Journal of Data Analysis and Information Processing*, *7*(4), 190–207. https://doi.org/10.4236/jdaip.2019.74012

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*(xxxx), 105361. https://doi.org/10.1016/j.knosys.2019.105361

Hemmer, P., Schemmer, M., Kuhl, N., Vossing, M., & Satzger, G. (2023). COMPLEMENTARITY IN HUMAN-AI COLLABORATION: CONCEPT, SOURCES, AND EVIDENCE. *Nature Medicine*, *29*(7), 1814–1820. https://doi.org/10.1038/s41591-023-02437-x

Hermiati, A. S., Herteno, R., Indriani, F., & Saragih, T. H. (2024). A Comparative Study : Application of Principal Component Analysis and Recursive Feature Elimination in Machine Learning for Stroke Prediction. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *6*(2), 231–242. https://doi.org/10.35882/jeeemi.v6i3.446

Li, B., Wu, Y., Zhang, Y., Hu, C., Li, X., Luo, S., Sun, C., & Yousef, I. (2025). Global and China trends and forecasts of disease burden for female lung Cancer from 1990 to 2021 : a study based on the global burden of disease 2021 database. *Journal of Cancer Research and Clinical Oncology*, *2*, 1–15. https://doi.org/10.1007/s00432-025-06084-2

Linardatos, P., & Papastefanopoulos, V. (2021). Explainable AI : A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 2–45. https://doi.org/10.3390/e23010018

Ma, W., Zhang, X., Shen, Y., Xie, J., Zuo, G., Zhang, X., & Jin, T. (2024). Incorporating Recursive Feature Elimination and Decomposed. *WATER*, *16*(21), 2–27. https://doi.org/10.3390/w16213102

Mani, K., & Rajaguru, H. (2024). Heliyon A framework for performance enhancement of classifiers in detection of prostate cancer from microarray gene. *Heliyon*, *10*(9), e29630. https://doi.org/10.1016/j.heliyon.2024.e29630

Nakhipova, V., Kerimbekov, Y., Umarova, Z., Suleimenova, L., & Botayeva, S. (2024). Use of the Naive Bayes Classifier Algorithm in Machine Learning for Student Performance Prediction. *International Journal of Information and Education Technology*, *14*(1), 92–98. https://doi.org/10.18178/ijiet.2024.14.1.2028

Parisi, F., Luca, G. De, Mosconi, M., Lastraioli, S., Dellepiane, C., Rossi, G., Puglisi, S., Bennicelli, E., Barletta, G., Zullo, L., Santamaria, S., Mora, M., Ballestrero, A., Montecucco, F., Bellodi, A., Del, L., Lambertini, M., Barisione, E., Cittadini, G., … Genova, C. (2024). Cancer Treatment and Research Communications Front-line liquid biopsy for early molecular assessment and treatment of hospitalized lung cancer patients. *Cancer Treatment and Research Communications*, *41*(August), 100839. https://doi.org/10.1016/j.ctarc.2024.100839

Richens, J. G., Lee, C. M., & Johri, S. (2020). with causal machine learning. *Nature Communications*, *11*(1), 1–9. https://doi.org/10.1038/s41467-020-17419-7

Simon, S. M., Glaum, P., & Valdovinos, F. S. (2023). Interpreting random forest analysis of ecological models to move from prediction to explanation. *Scientific Reports*, *0123456789*, 1–12. https://doi.org/10.1038/s41598-023-30313-8

Thabtah, F., Hammoud, S., & Kamalov, F. (2019). Data Imbalance in Classification : Experimental Evaluation. *Information Sciences*. https://doi.org/10.1016/j.ins.2019.11.004

Uddin, P., Mamun, A., & Hossain, A. (2020). PCA-based Feature Reduction for Hyperspectral Remote Sensing Image Classification PCA-based Feature Reduction for Hyperspectral Remote Sensing Image. *IETE Technical Review*, *0*(0), 1–21. https://doi.org/10.1080/02564602.2020.1740615