Lung Cancer Detection: AI Meets Healthcare

Zarin Saima ID: 24141186
, Lamisa Mahmud ID: 23241122
, Toufic Ahamad Piyeal ID: 23241121

[a]*BRAC University,*

Abstract

Lung cancer is the leading cause of cancer-related deaths worldwide, which makes it even more unanticipated that early and correct detection is vital. The present study is crucial to improving lung cancer prediction and diagnosis using machine learning (ML), specifically pattern recognition algorithms. We wish to construct a prediction model which can classify lung cancer instances, using a publicly available lung cancer dataset that contains patient records of various attributes. Therefore, the final part of the proposed research paper is to access more machine learning techniques (Logistic Regression, SVM, random forests, NN, Gradient Boost, XGBoost, AdaBoost, KNN, Decision Tree, Gaussian Naive Bayes Multinomial Naive Bayes) to determine which one is the best approach in classification. In the end, the effectiveness of these methods is evaluated by using evaluation metrics like accuracy, Precision, Recall, F1 Score, and Confusion Matrix of the models. These results emphasize the importance of machine learning-based pattern recognition in the early diagnosis of lung cancer, which may potentially influence clinical decisions and patient outcomes. This study shows that Gradient Boost achieves high performance, an accuracy of 95.39%.

Keywords: Lung cancer, Classification, Machine learning, Healthcare, Detection

1. Introduction

Millions of people die every year due to lung cancer and this is one of the leading causes of cancer-related deaths worldwide. Incredibly little progress has been made in arriving at an early diagnosis, perhaps explaining why the end result of all that research on lung cancer has been so dismal. For a better prognosis, accurate diagnosis of lung cancer is the key, and early-stage treatment success to improve survival. Traditional approaches like X-rays, biopsies and clinical evaluation have limited accuracy as well as time sensitivity for diagnosis. This highlights the need for more accurate and faster diagnostic methods. (1) During the past several years, one of the most transformative technologies in medical diagnostics-particularly for oncology-has been machine learning (ML). ML algorithms can identify patterns in big data, thus facilitating better predictive models and categorizing between benign-malign lung cancer cases. ML is appropriate, however, as it can quickly consume multiple patient records as well as various clinical data that greatly contribute to initial lung cancer diagnosis- an identity of prime importance for the clinical to aid in the information of his rationale.(2) The study aims to investigate a variety of machine learning (ML) algorithms to construct an accurate prediction model for lung cancer. The dataset used involves patient characteristics, and clinical history to lay a robust foundation for study. The classification in the study will be done using several machine learning (ML) techniques such as logistic Regression, Support Vector Machine (SVM), Random Forests, Neural Network, Gradient Boost, XGboost, AdaBoost, KNN, Decision Tree, Gaussian Naive Bayes Multinomial Naive Bayes.

We will assess the performance of these models using important metrics like accuracy, sensitivity, specificity, precision and F1 score. These metrics do not only indicate the general performance, but also potential reliability of the model when it comes to actual predictions. In comparative analysis, the level of screening model for early lung cancer detection is measured with utmost power hence increasing the diagnostic accuracy and possibly enhancing the full survival rates of patients in this study.

This article describes the methods used to tune the machine learning models, the performance metrics assessed and what our findings mean. This research aims to take full advantage of machine learning applications in diagnosing lung cancer and create the basis for more resilient diagnostic tools and better health outcomes.

2. Related Work

Recent advancements in ML techniques have significantly improved the prediction and early diagnosis of lung cancer. Various studies have explored different ML models and dataset to enhance the accuracy and performance of lung cancer detection. In one study, employed seven classification models such as decision trees, random forests, SVM, on a dataset of 15,750 clinical images, yielding the multilayer perceptron model as the most accurate at 88.55%, another investigation utilized a radial basis function network, achieving an accuracy of 81.25%(3).

Further contributions include a comparative analysis of multiple classifiers(4), where logistic regression demonstrates a remarkable accuracy of 96.9% on one dataset and 99.2% using SVM on another dataset. The integration of random forests with regression models in a study (5)successfully predicted patient survival time, showcasing the potential of hybrid models. In exploring early-stage lung cancer detection, researchers(6). The best performance was obtained by ensemble models, and the algorithm that employed the gradient-boosted trees had an accuracy of 90%. There is also another study where the application of data mining techniques is evaluated. In this research, the ann algorithm attained an accuracy of 92% while the CRISP-DM methodology was used (7). Biomarker identification to diagnose lung cancer at early stages was a crucial topic as well. It was established that this goal could be completed by mechanisms that included metabolomics and ML models since the obtained AUC was 0.989, which also means that the efficiency of the mechanism is highly linked with the appropriate selection of attributes for data processing procedures(8). Compared to the other studies, the present research uses a wide variety of classifiers, such as logistic regression, decision trees, and gradient boosting. The performance of these models is evaluated in regard to a special dataset on lung cancer in the context of feature engineering and pre-processing stages. Additionally, CV mechanisms are employed as a part of the evaluation procedure to discuss the advantages of a highly structured approach to model evaluation and assess the accuracy of predictions that are available to researchers due to the current approach, which may also help in addressing the identified problems of imbalance and lack of the model's interpretative capacity.

3. Literature Review

One of the major problems that is globally encountered includes lung cancer- Most of the people dying from lung cancer have made up a very huge portion of all cancer deaths every year. Lung cancer accounts for an estimated 1.8 million annual deaths — and is deemed by the World Health Organization (WHO) a public health priority. Among many reasons why the mortality rate from lung cancer is so devastatingly high: it is too often diagnosed at a late stage.(9) Lung cancer, particularly in the early stage, has little or no symptoms and could reduce the chances for earlier detection and proper treatment. In 2021, people 65 years of age and older accounted for 75% of lung cancer deaths, while people 55 years of age and older accounted for 96% of such fatalities. The death rate from lung cancer went up with age, reaching a peak of about 45,000 deaths in 2019 among people aged 65 to 74. After that, the death rate went down. In 2018, the death rate from lung cancer went up for people of all ages. The highest rate was found in people 85 and older, at 286.2 per 100,000..(10) For this reason, it is important to detect lung cancer as early as possible to increase survival rates and prevent patients from developing more advanced tumors.

4. Methodology

4.1. Dataset acquisition

This study uses a publicly available lung cancer dataset(11) in CSV format, composed of various features like patient demographics, medical history, and results. The dataset is analyzed to detect lung cancer. The dataset consists of 309 participants. The dataset includes 16 attributes in total, with 15 attributes as input features and 1 representing the target class.

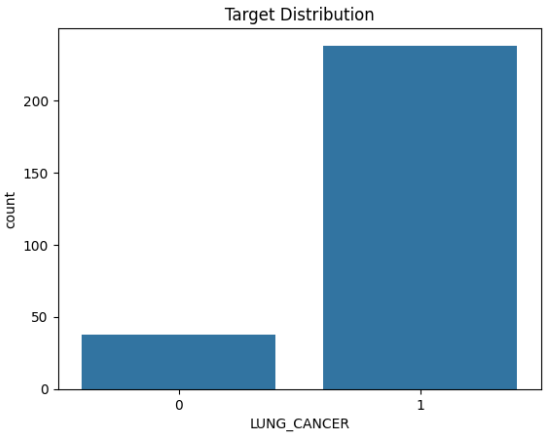

Figure 1: Distribution of Classes

4.2. Data Preprocessing

In the data preprocessing phase, the dataset was first checked for any null values, and duplicated rows were removed. Since the dataset contained numerous categorical features, these were converted to numerical values using appropriate encoding techniques. For attributes with only two unique categories, Label Encoding was applied. Additionally, as the age attribute had a wide range of values and was significantly larger compared to other numerical features in the dataset, Min-Max scaling was performed to normalize it. In the feature analysis phase, we examined Pearson's correlation using a heatmap. Since the features anxiety and yellow fingers exhibited the same effect on the dataset, these two features were combined. To address the highly imbalanced class distribution, where the Lung Cancer class accounted for 87.4% of participants, we employed the ADASYN model (12). ADASYN, a commonly used technique, utilizes a 5-NN classifier which creates synthetic samples for the minority class (Non-Lung Cancer). It is a process of oversampling that oversampled the non-cancer class, resulting in an equal distribution of both classes (i.e., 50%-50%). Lastly, to evaluate algorithm performance, we use the train test split method to divide the labeled data into two parts: 75% for training the model (training set) and 25% for testing its performance (test set). This allows us to see how well the model performs on new data.

4.3. Machine Learning Models

We applied different types of ML models to identify the most suitable classifier for lung cancer prediction, ensuring alignment with the research objectives.

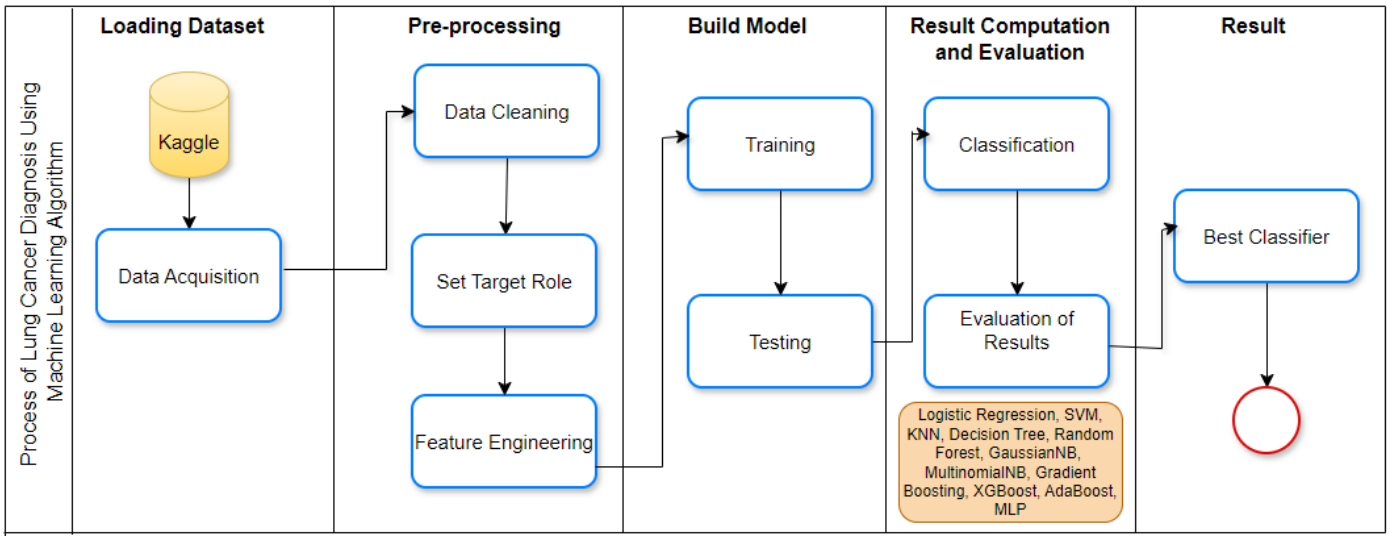

Figure 2: Process Flow Diagram

- Logistic Regression (LR) is a powerful modeling technique that generalizes linear regression(13). It assesses the probability of a disease or health condition based on risk factors and covariates. All types of logistic regression analyze the relationship between independent variables (Xi) and a binary dependent variable (Y), focusing on predicting binary or multiclass outcomes.
- Support Vector Machine (SVM) is a kernel-based classifier, to handle high-dimensional spaces effectively by finding a maximum marginal hyperplane (MMH) via the nearest data points (14).
- K-Nearest Neighbors (KNN) is one of the supervised classification algorithms that uses labeled data points to learn how to classify new, unlabeled points. To label a new point, the algorithm identifies its nearest neighbors—those labeled points closest to it—and assigns the most common label among them based on the voting system(15).
- Decision Trees (DT) is a predictive modeling tool applicable in various fields. It utilizes an algorithmic approach to divide the dataset in multiple ways based on different criteria(16). It gives a tree format to find the final prediction.
- Random Forest (RF) is also a tree-based model. It is one of the ensemble methods which also solve classification and regression problems. During the training phase they work by building a large number of decision trees and then aggregating their outputs—either by taking the mode of the prediction classes (for classification) or the mean prediction (for regression). This approach solves the problem of decision trees overfitting their training data.
- Naive Bayes (GaussianNB, BernoulliNB, MultinomialNB) to evaluate probabilistic models. They assume independence among predictors, making them computationally efficient for large datasets.(17)
- Gradient Boosting, XGBoost, and AdaBoost(18) for iterative boosting-based learning.
- Multi-Layer Perceptron (MLP) is a simple artificial neural network used for non-linear data analysis with neural networks.(19)

4.4. *Evaluation Metrics*

To check the performance evaluation of the machine learning models, metrics such as accuracy, precision, recall, F1-score, and AUC were taken into account (20). The required metrics will be calculated using the confusion matrix including true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

1. Accuracy:

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

2. Precision:

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

3. Recall (Sensitivity):

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

4. F1 Score:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where: TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives

Accuracy provides an overall summary of a classification task by measuring the proportion of correctly predicted instances out of the total data. Additionally, recall, which reflects the true positive rate, indicates how well the model identifies participants who actually had lung cancer and were correctly

classified as positive, in relation to all positive cases. While precision focuses on the quality of positive predictions, recall emphasizes the quantity of those correctly identified. The F1 score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives.Finally, the AUC ranges from zero to one and serves as a metric to identify the machine learning model that most effectively discriminates between lung cancer and non-lung cancer instances. AUC reflects the model's ability to separate the two classes. An AUC of one indicates that the model can perfectly distinguish between the two class distributions. Finally, in the cross-validation phase, we used Stratified K-Fold cross-validation to ensure that both the test set and the training set maintain a proportional representation of all categorical attributes. This approach helps confirm that the distribution of values across categories is preserved in both sets.

4.5. Ethical Considerations

All data were anonymized to protect patient privacy, and no personally identifiable information was utilized. The study aligns with ethical guidelines for medical data usage, ensuring that patient data is handled with care.

4.6. FAIR and CARE Principles

- FAIR (Findable, Accessible, Interoperable, Reusable): The dataset is openly available, ensuring reproducibility and transparency. The models, code, and results will be shared via open repositories for future research and clinical use.
- CARE (Collective Benefit, Authority to Control, Responsibility, Ethics): The research considers ethical data use, especially in the medical field. Models and data usage emphasize improving community health outcomes, ensuring that findings benefit society, not just specific stakeholders.

5. Results and Discussion

5.1. Experiment Environment

The scikit-learn library was used to do all research involving the machine learning models discussed in this study within the Python programming language. Sklearn is an open-source machine learning model. It has inbuilt models for classification, regression, and clustering, such as support vector machines, random forests, gradient boosting, k-means etc.(21). It is built to work seamlessly with Python's libraries, NumPy and Matplotlib.

5.2. Evaluation

In this study, plenty of Machine Learning models are evaluated in terms of accuracy, precision, recall, F1 score , confusion matrix adn AUC to determine the best predictive model.The following table shows the accuracy of testing and training set of the model. And SVC has the highest accuracy in terms of Testing accuracy.

Algorithms	Training Set (%)	Testing Set (%)
Logistic Regression	92.5	94.1
Random Forest	97.2	98.3
Decision Tree	97.2	93.3
Gaussian NB	88.5	90.8
Multinomial NB	79.6	84.2
SVC	96.4	99.2
Gradient Boost	96.9	96.7
XGBoost	96.9	98.4
MLP	96.1	96.7
K-NN	93.5	96.7
AdaBoost	92.7	95.8

Table 1: Accuracy percentage for Lung Cancer diagnostic dataset.

Algorithms	AUC(%)
Logistic Regression	99.4
Random Forest	99.6
Decision Tree	96.1
Gaussian NB	98.4
Multinomial NB	89.1
Gradient Boost	99.5
XGBoost	99.5
MLP	99.5
K-NN	98.7
AdaBoost	99.3

Table : Area Under ROC Curve

Algorithm	Precision	Recall	F1 Score	Class
LR	0.93	0.95	0.94	0
-	0.95	0.93	0.94	1
RF	0.98	0.98	0.98	0
-	0.98	0.98	0.98	1
DT	0.91	0.97	0.94	0
-	0.96	0.90	0.93	1
GaussianNB	0.95	0.87	0.90	0
-	0.88	0.95	0.91	1
MultinomialNB	0.85	0.83	0.84	0
-	0.84	0.85	0.84	1
SVC	0.98	1	0.99	0
-	1	0.98	0.99	1
GB	0.95	0.98	0.97	0
-	0.98	0.95	0.97	1
XGBoost	0.98	0.98	0.98	0
-	0.98	0.98	0.98	1
MLP	0.95	0.98	0.97	0
-	0.98	0.95	0.97	1
K-NN	0.94	1	0.97	0
-	1	0.93	0.97	1
AdaBoost	0.97	0.95	0.96	0
-	0.95	0.97	0.96	1

Table 2: Accuracy percentage for Lung Cancer diagnostic dataset.

In table 2, class 0 mean, no cancer, class 1 mean cancer detected.Among other models, SVC performs the best overall with an almost perfect F1 Score of 0.99 for both classes (0 and 1). This means it is highly effective in balancing precision and recall, leading to very few false positives and false negatives.

To confirm that we got the best modeland it performs well on all combination of train-test dataset we applied Stratified 10 fold cross validation. After that we get the following accuracy for all models-

Algorithms	Accuracy after CrossVal (%)
Logistic Regression	92.5
Random Forest	94.9
Decision Tree	93.3
Gaussian NB	87.8
Multinomial NB	78.8
SVC	93.7
Gradient Boost	95.4
XGBoost	94.5
MLP	93.7
K-NN	94.7
AdaBoost	92.5

Table 4: Accuracy after CrossVal

After applying 10-fold stratified cross-validation the best accuracy is the Gradient Boost model which has a testing accuracy of 95.39% which means before slightly the model over fitted. Finally, from this study we can tell that Gradient Boost worked as the best model.

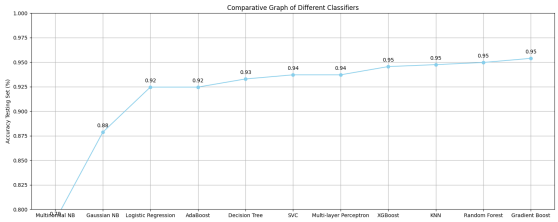

Figure 3: Accuracy After cross validation

5.3. Discussion

The methodology described in this work relies on a dataset comprising elements that reflect human behaviors, such as smoking and alcohol consumption, along with signs and symptoms that are typically associated with lung cancer patients as risk factors. Nevertheless, these indicators are not inherently associated with lung cancer, as evidenced by the feature analysis in the Materials and Methods section. In contrast to other malignancies, lung cancer is not visible to the naked eye, and its symptoms frequently coincide with those of other diseases. The predominant symptoms include allergies, asthma, dyspnea, and coughing(22). This study involved training many classifiers on diverse risk factors linked with certain symptoms to accurately determine the label of an unknown instance, along with the corresponding risk. Even in the absence of disease manifestation, monitoring risk factors and doing follow-up clinical examinations are prudent procedures in lung cancer management that may avert or mitigate the adverse effects of the disease through early identification. Results of this study show the advancement of ML in enhancing lung cancer detection, especially for early diagnosis, which is a key to higher survival rates. This work provides a wide comparison between the models made using SVM, Random Forests, NN, XGBoost, AdaBoost and KNN combined with multiple ML algorithms (Support Vector Machines-SVM, Random Forests-RF, Neural Networks-NN, XGBoost) in order to allows for different model selection. The main advantage of this study is the highlight on the clinical implementation of such machine learning models. The research combines sophisticated diagnostic approaches with lung cancer datasets from the real world and is thus able to bridge the gap between computational models and actual medical use. Using theAUC and Precision-Recall curves to evaluate model performance is important because it helps clinicians in understanding the trade-offs between sensitivity (catching every possible case) and specificity (avoid false alarms). That kind of advance could be revolutionary because early detection via ML models has the potential to transform the lung cancer diagnosis process from one predominated by traditional imaging and biopsy methods, to data-driven diagnostic tools. These models may be embedded in the healthcare system to support health providers in identifying patients at high risk and hence inform further testing or treatment options. Somewhat conversely, these algorithms might help to standardize how we diagnose things, mitigating the variability of human judgment and making initial cancer detection more accurate. In concluding the results and discussion part, it is essential to highlight an imperfection of our research. This study used a publicly accessible dataset (23), rather than data sourced from a medical center, which may provide more comprehensive information with diverse characteristics. Moreover, obtaining access to sensitive medical information is challenging due to privacy concerns. Nevertheless, the dataset upon which we depended possessed advantageous attributes that enabled us to obtain dependable and precise research outcomes.

6. Conclusion

The lungs serve as the primary organs responsible for respiration. Humans continuously breathe until death, as the lungs provide oxygen to the blood, which is essential for sustaining life. Lung cancer ranks as the foremost cause of mortality from malignancies in both males and females. The patient's lifespan is influenced by the advanced stage of cancer. Early diagnosis is associated with increased life expectancy. This study focuses on using supervised learning to create models for detecting individuals with lung cancer symptoms based on various features. A range of machine learning models, such as Logistic Regression, Decision Tree, KNN, Gaussian Naive Bayes, Multinomial Naive Bayes, SVC, Random Forest, XGBoost, Gradient Boost, Multi-layer Perceptron and AdaBoost were assessed based on accuracy, precision, recall, F1-score, and AUC. The research describes that stratified 10-fold cross-validation with implementing ADASYN, the Gradient boost

model demonstrated better performance compared to the rest of the models with an accuracy of 95.39%.

6.1. Future Work

We intend to expand the present investigation along three dimensions. By introducing the deep learning model's different CNN architecture we can work with image dataset. Lastly, would like to incorporate Quantum ML to detect lung cancer.

References

[1] Cleveland Clinic, "Lung cancer," May 1 2024, https://my.clevelandclinic.org/health/diseases/4375-lung-cancer (accessed on 2024-09-08). [Online]. Available: https://my.clevelandclinic.org/health/diseases/4375-lung-cancer

[2] C. Staff, "What is machine learning in health care?" Coursera, February 2024, available online: https://www.coursera.org/articles/machine-learning-in-health-care.

[3] G. A. P. Singh and P. Gupta, "Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6863–6877, 2019.

[4] P. Radhika, R. A. Nair, and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms," in *2019 IEEE international conference on electrical, computer and communication technologies (ICECCT)*. IEEE, 2019, pp. 1–4.

[5] J. A. Bartholomai and H. B. Frieboes, "Lung cancer survival prediction via machine learning regression, classification, and statistical techniques," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2018, pp. 632–637.

[6] M. I. Faisal, S. Bashir, Z. S. Khan, and F. H. Khan, "An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer," in *2018 3rd international conference on emerging trends in engineering, sciences and technology (ICEEST)*. IEEE, 2018, pp. 1–4.

[7] E. Vieira, D. Ferreira, C. Neto, A. Abelha, and J. Machado, "Data mining approach to classify cases of lung cancer," in *World Conference on Information Systems and Technologies*. Springer, 2021, pp. 511–521.

[8] Y. Xie, W.-Y. Meng, R.-Z. Li, Y.-W. Wang, X. Qian, C. Chan, Z.-F. Yu, X.-X. Fan, H.-D. Pan, C. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Translational oncology*, vol. 14, no. 1, p. 100907, 2021.

[9] J. Stern, J. Pier, and A. A. Litonjua, "Asthma epidemiology and risk factors," in *Seminars in immunopathology*, vol. 42, no. 1. Springer, 2020, pp. 5–15.

[10] W. H. Organization, "Lung cancer," World Health Organization, June 2023, available online: https://www.who.int/news-room/fact-sheets/detail/lung-cancer#:~:text=GLOBOCAN%202020%20estimates%20of%20cancer,deaths%20(18%25)%20in%202020.

[11] Sarah Mad Bhat, "Lung cancer prediction dataset," 2022, available online: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer (accessed on 3 July 2022).

[12] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Adasyn," Available online: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.ADASYN.html, 2024, accessed on: [Date you accessed the page].

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY: Springer-Verlag, 2001.

[14] W. S. Noble, "What is a support vector machine?" *Nature Biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[15] D. T. Larose, *Discovering Knowledge in Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2004.

[16] J. R. Quinlan, *C4.5: Programs for Machine Learning*, 2014, available online: https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1.

[17] D. Berrar, "Bayes' theorem and naive bayes classifier," 2019.

[18] K. Polat and U. Sentürk, "A novel ml approach to prediction of breast cancer: Combining of mad normalization, kmc based feature weighting and adaboostm1 classifier," in *2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. Ieee, 2018, pp. 1–4.

[19] N. Masih, H. Naz, and S. Ahuja, "Multilayer perceptron based deep neural network for early detection of coronary heart disease," *Health and Technology*, vol. 11, pp. 127–138, 2021.

[20] M. Zaman and C.-H. Lung, "Evaluation of machine learning techniques for network intrusion detection," in *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2018, pp. 1–5.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, V. Nicolas, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] L. A. Mandell and M. S. Niederman, "Aspiration pneumonia," *New England Journal of Medicine*, vol. 380, no. 7, pp. 651–663, 2019.

[23] J. A. Barta, C. A. Powell, and J. P. Wisnivesky, "Global epidemiology of lung cancer," *Annals of global health*, vol. 85, no. 1, 2019.