Lunch Cancer Detection Comparative Analysis using 7 Machine Learning Models

MD Alvin Sarkar Sakib
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
Email: 1903033@student.ruet.ac.bd

Shyla Afroge
Assistant Professor
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
Email: shyla.ruet@gmail.com

***Abstract*—Lung cancer is the top cause of death worldwide, necessitating the development of more effective diagnostic technologies for early identification. This study compares six machine learning models: support vector machine (SVM), logistic regression, Gaussian Naive Bayes, random forest, xgboost, and multilayer perceptron classifier with gradient boosting. Using cancer patient datasets, we assess how well these models diagnose lung cancer. Our study is to identify the most effective model for accurate and efficient lung cancer detection, which could lead to better patient outcomes and earlier intervention in this critical disease.**

***Index Terms*—Machine Learning Algorithms, Support Vector Machine, Lung Cancer Prediction, Gaussian Naive Bayes, Multilayer Perceptron classifier, XGBoost, Classification**

I. Introduction

Lung cancer throws a vast shadow throughout the globe, killing millions of people each year. Early identification is the most important aspect in successful therapy, providing a window of opportunity for curative measures. Traditional diagnostic approaches, while useful, can be time-consuming, subjective, and may miss early-stage malignancies. This needs the development of innovative, efficient, and objective ways to detecting lung cancer. According to the latest WHO data published in 2020 Lung Cancers Deaths in Bangladesh reached 12,174 or 1.70% of total deaths. The age adjusted Death Rate is 10.24 per 100,000 of population ranks Bangladesh #102 in the world[1]. Machine learning (ML) has emerged as a strong tool in medicine, with the potential to transform lung cancer diagnosis. ML systems can evaluate large amounts of medical data, such as chest X-ray pictures, and detect minor patterns that indicate malignant lesions. This ability to learn and adapt has enormous potential for early and precise lung cancer diagnosis.

This study investigates the potential of six different machine learning models: support vector machine (SVM), logistic regression, Gaussian Naive Bayes, random forest, xgboost, and multilayer perceptron classifier with gradient boosting. We assess and contrast the algorithms' performance in detecting lung cancer using datasets of different patients.

Our goal is to find the most effective and efficient ML model for accurate lung cancer detection. This study has important implications for the future of lung cancer diagnosis. By using the power of machine learning, we may be able to provide doctors with a faster, more objective tool for early identification, ultimately improving patient outcomes and delivering a ray of hope in the fight against this formidable disease.

II. Related Past Works

Previous research in lung cancer detection using machine learning (ML) techniques has extensively utilized text-based datasets to enhance prediction and diagnosis. For instance, the study [2] delves into various ML algorithms applied to text data extracted from medical records. The authors emphasize the utility of natural language processing (NLP) in transforming unstructured text into meaningful features, thereby improving the accuracy of lung cancer risk predictions. The [3] investigates the performance of multiple ML models in identifying lung cancer from text data. This comparative analysis underscores the significance of selecting appropriate algorithms and preprocessing techniques to optimize detection results. Further, the paper [4] explores advanced NLP models like SBERT and SimCSE on raw DNA sequences for cancer detection, including lung cancer. This study illustrates the potential of leveraging sentence transformers in combination with ML algorithms to enhance diagnostic accuracy. In addition, the work [5] demonstrates the value of clinical text data in predicting lung cancer prognosis. By analyzing clinical reports, the study shows how text data can provide insights into disease progression, which can inform tailored treatment plans. The research [6] focuses on utilizing ML and NLP techniques on electronic health records (EHRs) for early lung cancer detection. EHRs, rich in textual data such as patient histories and symptom descriptions, offer a valuable resource for early diagnostic efforts.

Moreover, the study [7] highlights the use of ML on medical text data to identify individuals at high risk of lung cancer, facilitating early intervention and screening. Lastly, [8] explores a hybrid model combining clinical text notes with chest X-ray images through deep learning techniques. This innovative approach exemplifies the potential of integrating multiple data sources to enhance the precision of lung cancer detection. These studies underscore the significant advancements in using ML and NLP techniques on text-based datasets to improve lung cancer detection and prognosis, demonstrating the transformative impact of these technologies in the medical field.

III. Dataset

The datasets used to identify lung cancer using machine learning models are rigorously constructed to include a wide range of patient features and risk factors. Each entry has an index, a unique patient identity, and demographic information like age and gender. Environmental exposures (for example, air pollution, alcohol consumption, dust allergy, and occupational risks) are recorded with genetic predispositions and pre-existing illnesses such as chronic lung disease. Lifestyle factors such as food, obesity, smoking habits, and passive smoking exposure are also documented. Clinical symptoms include chest pain, coughing up blood, exhaustion, weight loss, shortness of breath, wheezing, difficulty swallowing, fingernail clubbing, frequent colds, dry cough, and snoring. Each patient entry is eventually designated as a risk.

IV. Classification Algorithms

A. Logistic Regression

Logistic Regression is a linear classification model that estimates the probability of an event occurring based on a set of independent variables. It maps the linear combination of these variables (weighted sum) to a probability value between 0 and 1 using the sigmoid function. Logistic Regression minimizes a loss function (often the log loss) to find the optimal weights and bias that best separate the classes.

Let $\mathbf{x}$ be a vector of input features, $\mathbf{w}$ be the weight vector, and b be the bias term. The linear combination is $z = \mathbf{w}^T\mathbf{x} + b$. The sigmoid function, denoted by $\sigma(z)$, transforms this into a probability:

$$P(y = 1|\mathbf{x}) = \sigma(z) = \frac{1}{1 + \exp(-z)}$$

B. Gaussian Naive Bias

Gaussian Naive Bayes is a probabilistic classifier that assumes conditional independence between the features given the class label. It uses a Gaussian distribution (normal distribution) to model the continuous features for each class. It predicts the class with the highest posterior probability.

For a data point $\mathbf{x}$ with features $x_1, \ldots, x_d$ and class label y:

$$P(y|\mathbf{x}) \propto P(y) \prod_{i=1}^{d} P(x_i|y)$$

where: $P(y)$ is the prior probability of class y, $P(x_i|y)$ is the probability of feature x_i given class y, modeled as a Gaussian distribution:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_i^2(y)}} \exp\left(-\frac{(x_i - \mu_i(y))^2}{2\sigma_i^2(y)}\right)$$

Here, $\mu_i(y)$ is the mean of feature x_i in class y and $\sigma_i^2(y)$ is its variance.

C. Support Vector Machine

SVMs are kernel-based learning models that find a hyperplane (decision boundary) in high-dimensional space to maximize the margin between the classes. This margin is defined by the support vectors, which are the data points closest to the hyperplane.

D. Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees for improved prediction accuracy and robustness to overfitting. Each tree is trained on a random subset of features and data points (bootstrapping). Predictions are made by aggregating the individual tree predictions (e.g., majority vote for classification).

E. XGBoost

XGBoost (eXtreme Gradient Boosting) is a powerful tree-based ensemble learning method that leverages gradient boosting with regularization to improve performance. It builds decision trees sequentially, focusing on learning from the errors of previously built trees.

At each iteration m, XGBoost minimizes a regularized objective function:

$$Obj^{(m)} = \sum_{i} L(y_i, F_{m-1}(x_i) + f_m(x_i)) + \Omega(f_m)$$

F. Multi Layer perceptron Classifier

MLPs are artificial neural networks with multiple layers of interconnected nodes (neurons). They learn complex relationships between features and target variables through backpropagation, a training algorithm that adjusts the weights of connections based on the prediction errors.

$$h_j = \sigma\left(\sum_{i=1}^{d} w_{j.i} x_i + b_j\right)$$

activation function for hidden neurons

$$\hat{y} = \sigma'\left(\sum_{j=1}^{H} w_{o.j} h_j + b_o\right)$$

activation function for output neuron

G. Gradient Boosting

Gradient boosting is a general framework for building ensemble models by sequentially adding weak learners (models) that focus on improving the predictions of the previous ensemble. It leverages the concept of gradients (direction of steepest descent in the loss function) to guide the learning process.

V. Experimental Results

Our study compared seven machine learning models for lunch cancer detection. XGBoost achieved the highest accuracy (100%), Random Forest (100%) and Gradient Boost (100%). Logistic Regression, despite its simplicity, achieved a respectable accuracy of 86.6%. This suggests that simpler models might be suitable for initial screening, while more complex models like XGBoost could be used for further analysis. However, limitations like data size and hyperparameter tuning should be addressed in future work. Future research could explore feature engineering and external data validation to enhance the robustness and generalizability of these models.

Accuracy Result of 7 Machine Learning Models:
Logistic regression models accuracy: 0.866
Gaussian naive bayes models' average accuracy: 0.90
Support Vector Classifier models' average accuracy: 0.518
Random forest models' average accuracy: 1.0
XGBoost models' average accuracy: 1.0
Multi-layer perceptron models' average accuracy: 0.944
Gradient boost models' average accuracy: 1.0

VI. Conclusion

This study investigated how well seven categorization methods performed using [your dataset description]. The methods that were assessed included Gradient Boosting, Random Forest, XGBoost, Multi-Layer Perceptron (MLP) Classifier, Support Vector Machine (SVM), Logistic Regression, and Gaussian Naive Bayes. On the test set, all models had excellent accuracy, frequently above 90%.

Even while almost complete accuracy is ideal, it's crucial to take into account the research's limitations:

Small dataset size: Excessive accuracy on a tiny dataset may not translate well to new information. Future research with bigger and more varied datasets is required. Overfitting: If the models are memorizing the training data instead of learning generalizable patterns, the high accuracy could be a sign of overfitting. To lessen this, strategies like hyperparameter tuning and cross-validation might be used.

VII. Future Works

One crucial direction is to assess the models' generalizability by training and evaluating them on significantly larger and more diverse datasets. This will reveal how well the models perform on unseen data, reflecting their real-world applicability. Additionally, incorporating domain-specific knowledge into the models can be highly beneficial. By leveraging expert insights and industry-specific information, researchers can potentially improve model performance and enhance interpretability. This can involve feature engineering techniques that tailor the data representation to the problem domain.

Furthermore, exploring alternative evaluation metrics beyond accuracy can provide a richer picture of the models' strengths and weaknesses. Metrics like precision, recall, F1-score, and AUC offer valuable insights into how well the models handle specific classification tasks. For instance, if the cost of misclassifying a particular class is high, precision or F1-score might be more relevant metrics than overall accuracy. Conducting a more rigorous hyperparameter tuning process for each model is also vital. Hyperparameters significantly influence model behavior, and a systematic approach to tuning them can lead to substantial performance improvements.

The future of classification research holds exciting possibilities in the realm of ensemble methods. These techniques combine multiple models, potentially achieving superior results compared to individual models. Researchers can explore various ensemble strategies, such as bagging, boosting, and stacking, to identify the best approach for the specific classification task. Finally, applying Explainable AI (XAI) techniques is crucial, especially for black-box models like MLPs and XGBoost. XAI helps us understand the decision-making processes behind these models, fostering trust and enabling further performance optimization. By contributing to the development of new XAI methods and applying them to classification tasks, researchers can significantly enhance the transparency and interpretability of these powerful models.

References

[1] W. H. O. 2020, "https://www.worldlifeexpectancy.com/bangladesh-lungcancers: :text=according

[2] K. Mohan and B. Thayyil, "Machine learning techniques for lung cancer risk prediction using text dataset," 2023.

[3] R. P. R, "A comparative study of lung cancer detection using machine learning algorithms," 2022.

[4] T. L. et al, "A review and comparative study of cancer detection using machine learning: Sbert and simcse application," 2023.

[5] J. C. et al., "Extracting actionable insights from clinical text reports for lung cancer prognosis," 2021.

[6] S. G. et al., "Leveraging machine learning and natural language processing for early lung cancer detection from ehrs," 2020.

[7] Z. Xu, "Text-based lung cancer risk assessment using machine learning," 2019.

[8] Y. Yang, "A hybrid deep learning approach for lung cancer detection using clinical text data and chest x-rays," 2018.