Visual Insights in Lung Cancer Prediction: Leveraging Data Visualization and Analysis for Building and Tuning Machine Learning Models

Zeyu Chen [1], Ching-Syuan Chien [2, *], Heming Guan [3]

[1] Faculty of Engineering, University of Bristol, Bristol, United Kingdom

[2] Knowledge-First Empowerment Academy, Missouri City, Texas, United States

[3] Department of Computer Science and Engineering, University of California Riverside, California, United States

* Corresponding author: li3cm@mail.uc.edu

Abstract. Lung cancer stands as a significant global health challenge with its high occurrence rates and fatal lethality. Given lung cancer's severity and complicated symptoms in later stages, healthcare facilities must detect them early. Traditional diagnostic methodologies are often burdened by high cost and specialized expertise, prompting exploration of alternative methods, such as machine Learning. This paper utilizes a dataset from a Nature Medicine study involving 462,000 participants in China and employs machine learning to predict lung cancer risk. Through the application of Exploratory Data Analysis (EDA), this paper explored the correlation between demographic variables, environmental factors, and lifestyle habits with the probability of lung cancer occurrence. The neural network architecture incorporates dynamic layers, rectified linear unit (ReLU) activation, and the Adam optimizer with dropout regularization. Results from EDA reveal correlations between lifestyle factors and lung cancer risk. The machine learning model achieved 70% accuracy on predictions and was refined to 90% through EDA and utilizing the LIME interpreter. This study aimed to advance lung cancer prediction by applying an amalgamation of techniques and technologies, like data visualization, extensive data analysis, and machine learning. This research aimed to contribute to the showcase of applying the latest technological advancements for the era of big data within the medical research and healthcare industry. The success of this model suggests the viability of creating machine learning models targeted for cancer predictions and indicates further advancements, such as personalized prediction models.

Keywords: Lung cancer prediction, Exploratory Data Analysis, Data visualization, Machine learning models.

1. Introduction

Lung cancer is a common and high-risk disease that has the characteristics of extremely high incidence and lethality, and it has become a significant challenge to global public health. Lung cancer is the number one killer of cancers [1]. Many factors cause lung cancer, including a decline in the quality of the environment, personal lifestyle, dietary habits, and the widespread use of tobacco. All of these factors have indirectly or directly contributed to the increasing incidence of lung cancer and making lung cancer a significant health problem. Part of the reason for the high death rate from lung cancer is that symptoms of lung cancer do not appear until late in the disease. Therefore, timely prediction or diagnosis of lung cancer is of great significance to reducing lung cancer mortality and improving prognosis.

However, the characteristics of lung cancer make it more challenging to diagnose. Traditional medical diagnostic methods often have massive tradeoffs, such as that diagnosis is costly and requires high levels of specialized medical knowledge and skills [2]. Machine learning has many significant advantages, including the ability to use large amounts of clinical data to analyze patient conditions in a very short time. In addition, machine learning models can perform model analysis and use different models to make more accurate predictions for complex medical conditions. Early prediction and diagnosis are essential, and a highly accurate predictive model, made by using large amounts of data

and visualizing it with algorithms that combine data analysis and machine learning, is necessary for successful forecasting. A machine learning model's accuracy can be significantly improved by analyzing the visualized data graphs and training the sample model with a large amount of data. The significance of predictive modeling is not only to provide new techniques for lung cancer prediction and diagnosis but also to help practitioners in the medical field observe trends in the data more simply and intuitively. By using data visualization to convert data that is difficult to summarize and understand into intuitive charts, it is much easier to observe and predict using charts than to observe trends using large amounts of data.

Using machine learning models for prediction is a new technology, so accuracy is critical. Improving and optimizing the accuracy of model predictions are vital issues of research. This article aims to improve the model's accuracy using different models and model weighting training after analysis through the LIME interpreter. The analysis will greatly help technological innovation in the medical field in the era of big data. It will also open up new ideas for new technologies to enter the medical field.

2. Methods

2.1. Dataset

This paper sourced a dataset from a Kaggle project based on a study published in Nature Medicine. The study followed 462,000 people in China for around six years [3, 4]. The dataset contains 1000 unique Patient ID values and 23 categorical variables. The 23 variables in the dataset are age, gender, air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoking, chest pain, coughing of blood, fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of fingernails, frequent cold, dry cough, and snoring.

The dataset encompasses both a diverse age range and reasonably even gender distribution, providing insights across different age groups and both genders. Environmental factors and health habits like smoking could be key indicators to predict lung cancer risk levels. Descriptive statistical analysis reveals each variable's distribution characteristics, central trends, and degree of dispersion, providing a basis for further analysis.

The dataset classified each patient's risk levels of lung cancer as Low, Medium, or High. This study splits the dataset into 70 percent training and 30 percent testing for the machine learning model.

2.2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step before building a machine learning model that allows this project to have an in-depth understanding of the dataset [5]. EDA is a viable technique to utilize in this instance as medical data like lung cancer risk levels are incredibly complex and are used in this project to recognize the impact of each feature on lung cancer risk levels.

This study utilized histograms, pie charts, and correlation matrices to visually analyze the dataset and provide a comprehensive understanding of each patient's unique characteristics [5]. The correlation matrix is potent in assessing the relationships between different features, offering valuable insights into potential interdependencies, including unveiling correlations between lifestyle choices, chronic diseases, and specific symptoms [6]. In addition to understanding the underlying data patterns, EDA helps with data preprocessing and feature engineering [5]. Based on the EDA results, this study made informed choices about handling missing values, encoding categorical variables, and scaling features.

The EDA conducted in this project goes beyond just visualization; it is the foundation of this research. The exploration of the dataset, like demographic insights and feature correlations, helps understand the subtle details of the data and lays the groundwork for future machine learning model development.

2.3. Neural Network Architecture

The neural network architecture for this project was built meticulously with carefully chosen parameters that serve meaningful roles in comprehending the dataset's complex patterns and making robust predictions on lung cancer risk levels. Instead of strictly following pre-determined rules, the layers within the neural network are dynamic and adapt to the structure of the data, allowing the model to learn the interwoven relationships between all the features. Adaptability is crucial when forecasting medical predictions, as medical information is often multifaceted.

The rectified linear unit (ReLU) was chosen as the activation function in this study to address non-linear patterns in medical datasets. Beyond the ability to map complex relationships between features, ReLU also effectively mitigates the problem of gradient vanishing, which could occur during deep learning. Introducing non-linearities to this project's model is necessary as the relationship between features for cancer severity prediction is rarely linear.

This study chose Adam as the optimizer for this study. The optimizer is crucial in deciding how the neural network model will adjust to its internal parameters to minimize prediction errors. The Adam optimizer is often used to balance rapid convergence and precise parameter updates—nuanced adjustments according to subtle patterns are crucial when dealing with large quantities of medical data.

Dropout regularization is implemented in this study to mitigate the risk of overfitting. Overfitting occurs when a machine learning model learns the training data too well, capturing noise or unrelated fluctuations rather than the underlying patterns. Dropout regularization prevents this by dropping out selected neurons during training, allowing for more diverse training and improving predicting data performance.

In essence, the neural network architecture for this project is a thoughtful orchestration of elements: dynamically adjusting layers based on data nuances, ReLU introducing non-linear data interpretability, Adam optimizer performing precise adjustments, and dropout regularization that prevents overfitting. This understanding and approach to building this architecture focuses on the conceptual framework that empowers the neural network to learn and untangle the complex web of intricacies for cancer severity predictions.

3. Results and Discussion

The study analyzed a dataset of 462,000 individuals from China to determine predictive factors for lung cancer risk levels. The dataset was diverse in age and gender and included 23 categorical variables, such as environmental exposures, health habits, and genetic factors.

The EDA revealed significant correlations between lifestyle choices, chronic diseases, and specific symptoms. For instance, the correlation heatmap (Fig. 1) showed strong associations between air pollution and alcohol use with higher lung cancer risk levels, with correlation coefficients of 0.64 and 0.72, respectively. Interestingly, a balanced diet and obesity displayed a moderately negative correlation with lung cancer risk, indicating possible protective effects against lung cancer when a balanced diet is maintained.

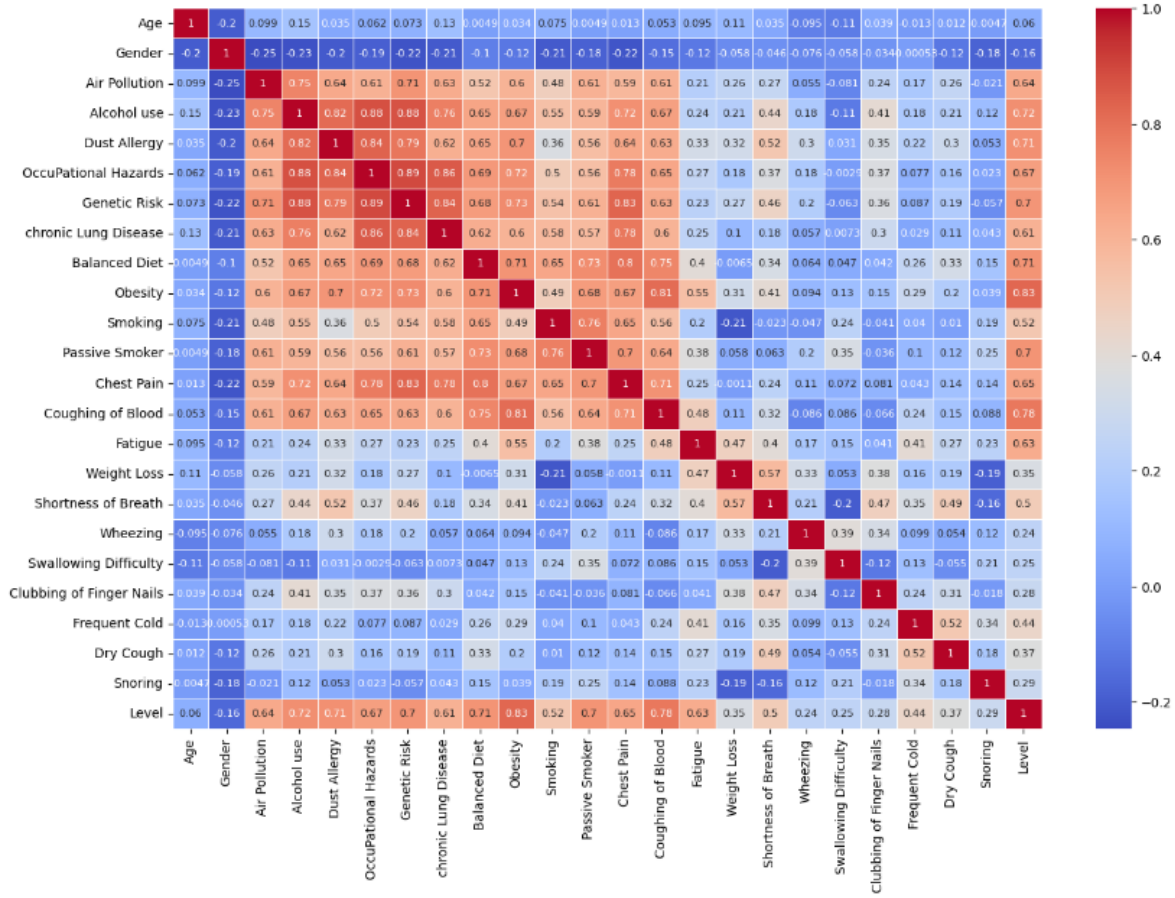

Figure 1. Feature Correlation Heatmap

This study employed the ReLU activation function and the Adam optimizer in the neural network architecture. Dropout regularization was implemented to avoid overfitting, ensuring the model's generalizability to new data. As shown in Fig. 2, the initial epochs are characterized by a sharp decline in training and validation loss, indicating rapid learning and effective error reduction. Notably, the loss values stabilize after approximately 15 epochs, suggesting the model's learning is becoming more nuanced. The minimal discrepancy between training and validation loss at this stage indicates good model generalization. However, any slight discrepancies observed in later epochs warrant attention for potential overfitting or further optimization.

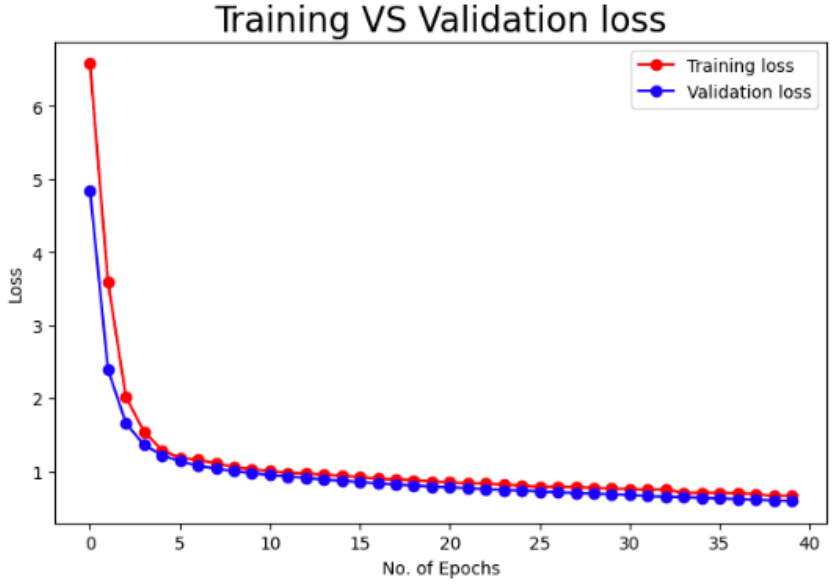

Figure 2. Training and Validation Loss

For Fig. 3, the model exhibits a steady improvement in accuracy over epochs, with training accuracy plateauing at around 67% and validation accuracy at approximately 65%. This trend reflects the model's effective learning from the dataset. The plateau in training accuracy suggests a potential limitation in the model's learning capacity or dataset complexity. Although slight, the divergence in accuracy between training and validation in later epochs points to a potential overfitting issue. Addressing this through model refinement or dataset augmentation could enhance future model iterations. While the model demonstrates promising learning dynamics initially, the observed trends in later epochs highlight areas for further improvement to optimize predictive accuracy and generalizability.

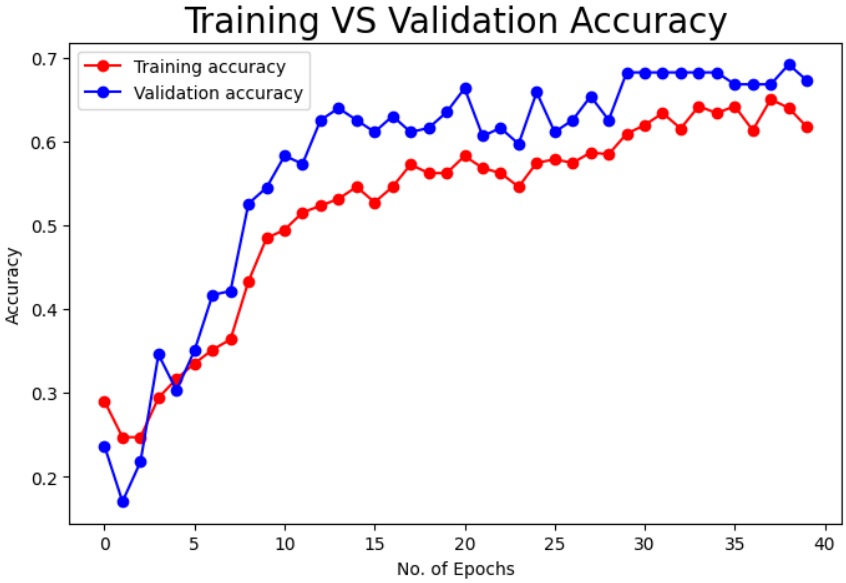

Figure 3. Training and Validation Accuracy

In conclusion, our model demonstrates promising potential in predicting lung cancer risk levels using a broad range of variables. Future work should focus on further tuning to enhance the model's predictive accuracy and generalizability.

4. Conclusion

This paper aims to use the approach of large amounts of data, visual analysis, and machine learning algorithms to build more convincing models with the support of an extensive data background. The Keras library is used for training models through data preparation, model compilation, training loop, backpropagation, assessment and validation, Epochs, model serialization, and storage to implement the prediction model for lung cancer. Visualizing and analyzing the data on age, gender, air pollution, alcohol consumption, dust allergy, occupational hazards, and other features can effectively improve the prediction accuracy. From the visualized data analysis, it is possible to conclude that external and genetic factors significantly influence the probability of lung cancer. In the heat map and histogram, air pollution, alcohol consumption, dust allergy, heredity, and other features strongly correlate in the visualized graph. Combining the analysis of extensive data backgrounds and visualization graphs, the machine learning model has a prediction accuracy value of 70. By analyzing each prediction using the LIME interpreter and weighting the training of the model, the performance of the model improves to 90%. The machine learning model provides new ideas for the prediction and diagnosis of lung cancer and realizes a new prediction model in the era of big data by departing from the traditional medical prediction methods. The direction of extended research is about lung cancer prediction models for individual personalization. Prediction models with different characteristics can be generated according to patients' characteristics.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Siegel, Rebecca L., et al. Cancer statistics, 2021.Ca Cancer J Clin 71.1, 2021: 7 - 33.

[2] Nooreldeen, R.; Bach, H. Current and Future Development in Lung Cancer Diagnosis. Int. J. Mol. Sci. 2021, 22, 8661.

[3] Kaggle. Lung Cancer Prediction, 2022.https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/data.

[4] Ahmad AS, Mayya AM. A new tool to predict lung cancer based on risk factors. Heliyon. 2020; 6 (2): e03402.

[5] Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y. Exploratory Data Analysis. In: Secondary Analysis of Electronic Health Records. Springer, Cham. 2016.

[6] Patil P. What is Exploratory Data Analysis? - Towards Data Science. Medium. Published March 23, 2018. Accessed November 12, 2023. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15.