**Lung Disease Detection using Convolutional Neural Network with the Implementation of XGBoost**

2602189953 - Austin Natan The

2602108975 - Karldritz Farrel Hanson

2602140616 - Marco Thierry Sondakh

1. **Introduction**

Diseases have existed and lived alongside humanity since the beginning of civilization. With the field of biological studies, humans have identified many diseases that ranged between minor inconveniences to life-threatening. One such life-threatening disease is lung disease. There are many forms of lung disease, each can either be easily treated or quite difficult to treat. The problem that comes with lung disease is that they display the same symptoms as a variety of health issues. This makes lung disease difficult to detect manually, which can lead to improper treatment in the early stages of a patient. This is where bioinformatics comes into play.

Studies on biology were created to not only learn about the fundamentals of life, but to also improve them. Bioinformatics is a branch of this study that combines biology with informatics to enhance the field using technologies that have advanced throughout the years. Thanks to bioinformatics, biological analysis and experimentation are made more efficient and cost-effective. Analysis of cells, DNA, and structures of living creatures are processed much faster with state-of-the-art technology, it is also time saving as a computer can process large quantities of information faster than what an average human can achieve. One field that is also improved using bioinformatics is disease detection.

With bioinformatics, we plan on creating a robust lung disease detection model using a deep learning algorithm, Convolutional Neural Network (CNN), which can robustly handle images and x-ray scans, and boost it using an ensemble learning model, XGBoost, which is often used in medical diagnostics task, usually to enhance the detection process of the diseases.

1. **Methodology**

Figure (3.1) . Training Pipeline

A black and white rectangular sign

Description automatically generated

1. **Dataset**

For our study, we utilized the lung disease images dataset available from the Kaggle repository: Lungs Disease Dataset (4 Types). This dataset comprises X-ray images categorized into four types of lung diseases:

|  |  |
| --- | --- |
| No | Lung Disease Type |
| 0 | Covid-19 |
| 1 | Normal |
| 2 | Viral Pneumonia |
| 3 | Bacterial Pneumonia |

Table(3.1). Types of Lung Disease In The Dataset.

The images in the dataset vary in size, and each type of lung disease is given labels accordingly to facilitate the training of our model. This diversity in the dataset is crucial for training a robust model capable of distinguishing between the different types of lung diseases.

1. **CNN**

We will be utilizing a novel Convolutional Neural Network (CNN), choosing from the InceptionV3, DenseNet, ResNet, or VGG, as the base for our model, with the top layer of the model being manually trained by ourselves. Additional parameters and tuning will be done to the model such as adding early stopping and learning rate reduction to help ensure the model becomes robust and does not tend to overfit.

With data preprocessing, the dataset is processed to fit the model which will allow for proper processing in the model. This model’s accuracy is compared to see which CNN model can show the best result for our research target.

1. **XGBoost**

XGBoost is a powerful gradient boost algorithm that will help us classify lung diseases based on the extracted feature by CNN. XGBoost is an ensemble technique that has multiple decision trees and combines their output to improve the accuracy and performance of the model. Moreover, XGBoost helps to capture or read complex patterns of images which can further improve lung disease classification accuracy. In the training of XGBoost features obtained from CNN will be split for training, testing, and validation dataset. Hyperparameter tuning will also be included in the model such as grid search alongside cross validation to avoid overfitting while improving the model further.

1. **Accuracy**

Accuracy is a widely used metric for evaluating the performance of classification models. It measures how well a model can correctly classify data overall. In our context, a "correct" prediction means that the model's prediction matches the true label of the observed data. The accuracy is calculated by dividing the number of correct predictions by the total number of predictions made by the model. The formula for accuracy is:

1. **Precision**

Precision is a crucial metric when needing to classify things, especially when minimizing false positives. It measures the accuracy of positive predictions made by the model, indicating the amount of correctly predicted positive instances out of all instances predicted as positive. The formula for precision is:

1. **Recall Score**

Recall, also known as sensitivity or the true positive rate, is essential when needing to identify all positive instances. It measures the model's ability to correctly determine positive instances from all actual positive instances in the dataset. The formula for recall is:

1. **F1-Score**

The F1 score combines precision and recall into a single metric, providing a balanced measure that is useful when both false positives and false negatives need to be minimized. The F1 score is calculated using the harmonic mean of precision and recall. The formula for the F1 score is:

The F1 score ranges from 0 to 1, with a higher score indicating better performance. It is particularly useful in evaluating classification models in scenarios where a balance between precision and recall is important, such as in medical diagnosis.

1. **Experiments and Results**

This section will revolve around explanations regarding our experiment and the results yielded from it. Experimentation took place in visual studio code utilizing the python programming language with the dataset mentioned previously. Discussions will be held to show the discrepancies in accuracy of each distinct model.

In our ensemble model, we are combining the features extracted by using Convolutional Neural Network models, and then an XGBoost model will be utilized to make proper classification. This approach takes the strengths of both CNN and XGBoost to enhance the accuracy and robustness of lung disease detection. Below the performance of the model highlighting its accuracy, precision, recall score, and f1 score can be seen.

Table (4.1) . Evaluation Result of DenseNet121 + XGBoost

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **0** | **1** | **2** | **3** | **4** |
| Precision | 0.72 | 0.86 | 0.81 | 0.92 | 0.64 |
| Recall | 0.66 | 0.89 | 0.91 | 0.92 | 0.60 |
| F1-Score | 0.69 | 0.88 | 0.86 | 0.92 | 0.62 |
| Accuracy | 0.80 | | | | |

Figure(4.1). Confusion Matrix of DENSENET121 + XGBOOST

A blue squares with numbers

Description automatically generated

Table (4.2) . Evaluation Result of VGG16 + XGBoost

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **0** | **1** | **2** | **3** | **4** |
| Precision | 0.66 | 0.75 | 0.79 | 0.89 | 0.57 |
| Recall | 0.53 | 0.85 | 0.86 | 0.88 | 0.56 |
| F1-Score | 0.59 | 0.80 | 0.82 | 0.89 | 0.56 |
| Accuracy | 0.74 | | | | |

Figure(4.2). Confusion Matrix OF VGG16 + XGBOOST

A blue squares with white text

Description automatically generated

Table (4.3) . Evaluation Result of ResNet50 + XGBoost

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **0** | **1** | **2** | **3** | **4** |
| Precision | 0.56 | 0.71 | 0.71 | 0.73 | 0.52 |
| Recall | 0.53 | 0.76 | 0.84 | 0.75 | 0.40 |
| F1-Score | 0.54 | 0.73 | 0.77 | 0.74 | 0.45 |
| Accuracy | 0.66 | | | | |

Figure(4.3). Confusion Matrix of RESNET50 + XGBOOST

A blue squares with white text

Description automatically generated

Table (4.4) . Evaluation Result of InceptionV3 + XGBoost

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **0** | **1** | **2** | **3** | **4** |
| Precision | 0.62 | 0.73 | 0.72 | 0.82 | 0.52 |
| Recall | 0.59 | 0.85 | 0.73 | 0.80 | 0.45 |
| F1-Score | 0.60 | 0.78 | 0.72 | 0.81 | 0.48 |
| Accuracy | 0.69 | | | | |

Figure(4.4). Confusion Matrix OF INCEPTIONV3 + XGBOOST

A blue squares with white text

Description automatically generated

The results indicate that the DenseNet121 + XGBoost model outperformed the other combinations, achieving the highest accuracy of 80%. Each model's performance varied in terms of precision, recall, and F1 score, highlighting the trade-offs between these metrics. The choice of the base CNN model significantly impacted the overall performance of the ensemble model, with DenseNet121 demonstrating superior capability in feature extraction for lung disease classification. While our results are comparable to the previously mentioned related works, the use of an ensemble method in our approach potentially enhances the stability and accuracy of lung disease detection. Our research contributes to the ongoing advancements in medical image analysis, providing a framework that integrates multiple models to achieve reliable and high-performance diagnostic outcomes.

1. **Conclusion**

This study aimed to develop a robust lung disease detection model by integrating Convolutional Neural Networks (CNNs) with XGBoost. The goal was to leverage the strengths of both CNNs in handling image data and XGBoost in enhancing classification performance. We tested various CNN architectures, including DenseNet121, VGG16, ResNet50, and InceptionV3, to determine the most effective model for extracting features from lung disease X-ray images.

The DenseNet121 + XGBoost model achieved the highest accuracy at 80%, outperforming the other CNN-based models. The metrics for precision, recall, and F1 score across different models illustrated the trade-offs and highlighted the superior performance of DenseNet121 in feature extraction. This ensemble approach, combining CNN and XGBoost, demonstrated a notable improvement in stability and accuracy for lung disease detection.

Our research was done to contribute to the field of medical image analysis by presenting a framework that integrates multiple models to enhance diagnostic accuracy. The findings suggest that using ensemble methods can provide reliable and high-performance outcomes in lung disease classification, thus potentially aiding in early detection and proper treatment of patients. Future work may explore further fine-tuning of the models and incorporating additional data to enhance the robustness and generalizability of the detection system

1. **Code Link**

[**https://github.com/Austinatan22/Machine-Learning-AOL**](https://github.com/Austinatan22/Machine-Learning-AOL)

1. **Datasets**

[**https://www.kaggle.com/datasets/omkarmanohardalvi/lungs-disease-dataset-4-types**](https://www.kaggle.com/datasets/omkarmanohardalvi/lungs-disease-dataset-4-types)