

FINAL PROJECT

OBJECTIVES OF THE PROJECT

The objectives of this project is to create a robust predictive for predicting pneumonia based on chest x-rays.

SIGNIFICANCE OF THE PROJECT

The creation of a machine learning algorithm to forecast pneumonia based on chest X-ray images is of great importance in the field of healthcare. Pneumonia continues to be a widespread and potentially fatal illness that has a significant impact on vulnerable groups like children, senior citizens, and those with compromised immune systems. It is crucial to have a prompt and precise diagnosis in order to intervene in a timely manner, lower mortality rates, and alleviate the strain on healthcare resources.

Utilizing machine learning algorithms that have been trained on chest X-ray images, this model has the ability to rapidly and precisely identify cases of pneumonia. This can greatly assist healthcare professionals in making prompt and well-informed decisions. The early detection of pneumonia through automated systems can significantly speed up the initiation of treatment, potentially averting complications and reducing the length of hospital stays. Furthermore, in regions where access to specialized healthcare is limited, this technology could prove to be an invaluable diagnostic tool, providing initial assessments that complement the efforts of healthcare providers.

Furthermore, the model's capacity to examine extensive quantities of imaging data can play a significant role in establishing recurring trends and detecting subtle cues of pneumonia that might escape the attention of human observers. The incorporation of this advanced technology into clinical workflows holds promising possibilities for bolstering diagnostic precision, streamlining healthcare resource distribution, and ultimately enhancing patient outcomes. By doing so, this project showcases its remarkable potential to revolutionize the entire process of pneumonia diagnosis and treatment, highlighting its profound impact on the field.

DATA

The data were downloaded from kaggle.com. The dataset is divided into three main folders, namely train, test, and val. Inside each of these folders, there are subfolders representing different image categories, namely Pneumonia and Normal. The dataset consists of a total of 5,863 X-Ray images in JPEG format, and these images are classified into two categories, Pneumonia and Normal.

Chest X-ray images captured in the anterior-posterior view were carefully chosen from a collection of previous cases involving young children aged between one and five years at the Guangzhou Women and Children's Medical Center in Guangzhou. It is important to note that these chest X-rays were not taken specifically for the purpose of this study, but rather as a routine part of the patients' regular clinical care.

In order to ensure accurate analysis of chest x-ray images, a thorough quality control process was implemented. Initially, all chest radiographs underwent a screening process to eliminate any scans of low quality or those that were unreadable. Once this initial screening was completed, the images were then assessed by two highly skilled physicians who graded the diagnoses. Only after this rigorous evaluation were the images deemed suitable for training the AI system. Additionally, to minimize the risk of any potential grading errors, a third expert carefully examined the evaluation set.

DATA EXPLORATION

This project removed dodgy images from the dataset by ensuring that, all the images have one of the following extensions; 'jpeg', 'jpg', 'bmp', 'png'. Further, all images less than 10kb were removed

CLASSIFICATION MODELS

This project used three (3) models. As seen in the table below, the CNN is the best performing model in terms of accuracy, precision, F-score and Recall. Therefore I selected CNN for this project.

MODELS	ACCURACY	PRECISION	F_SCORE	RECALL
CNN	0.95	0.92	0.95	0.98
VGG16	0.867925	0.842105	0.820513	0.8
GLOBALAVERAGEPOOLING2D	0.811321	0.75	0.75	0.75

KEY FINDINGS

The provided results are from a Convolutional Neural Networks (CNN) deep learning model, and they include various performance metrics, including accuracy, precision, F-score, and recall. These metrics are used to evaluate the model's performance in classifying whether a given chest X-ray is associated with pneumonia or normal

1. Accuracy: Accuracy is a measure of the overall correct classification rate of the model. In this case, the CNN model has an accuracy of approximately 0.95, which means it correctly classifies around 95% of the samples. This is a good overall accuracy.
2. Precision: Precision measures the ability of the model to make correct positive predictions among all positive predictions made. The precision of approximately 0.92 indicates that when the model predicts that an X-ray is associated with pneumonia, it is correct about 92% of the time.
3. F-Score: The F-score is the harmonic mean of precision and recall. It provides a balance between these two metrics. An F-score of approximately 0.95 is a good score and indicates that the model achieves a good trade-off between precision and recall.
4. Recall: Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify all actual positive cases. The recall value of 0.98 means that the model correctly identifies 98% of the X-rays which are actually associated with Pneumonia

In summary, the CNN model shows promising results for predicting pneumonia based on chest x-rays. It has a high accuracy, indicating that it correctly predicts the majority of cases. The precision and recall values are also relatively high, suggesting

that the model is good at correctly identifying pneumonia chest x-ray and making positive predictions when necessary.

LIMITATION OF THE MODEL AND FUTURE REVISION

Limited Model Comparison: The study examined only three models (CNN, GlobalAveragePooling2D, VGG16). Other architectures or variations were not explored, which might have led to potential better-performing models that were overlooked. Therefore, I intend to explore more sophisticated architectures and optimization techniques.