

***A Mini-Project Report On***

**“Pneumonia Detection ”**

***Submitted By***

**Abhishek Saha**

**1132220532**

**F.Y. M.Sc. (Data Science and Big Data Analytics)**

**School of Computer Science**

**Faculty of Engineering and Technology**

**Dr. Vishwanath Karad MIT – World Peace University**

**Pune - 411038**

**Academic Year 2022-2023**

**Nov - 2022**

**Dr. Vishwanath Karad MIT WORLD PEACE UNIVERSITY, PUNE**

**SCHOOL OF COMPUTER SCIENCE**

Certificate

This is to certify that

*Abhishek Saha*

*1132220532*

**Of M.Sc. (Data Science and Big Data Analytics) successfully completed his/her**

**Mini-Project in**

“Pneumonia Detection Project”

to our satisfaction and submitted the same during the academic year 2021 - 2022 towards the partial fulfilment of degree of Master of Science in Data Science and Big Data Analytics of Dr Vishwanath Karad MIT World Peace University under the School of Computer Science, MIT WPU, Pune.

Prof. Dr. Shubhalaxmi Joshi Prof. Surabhi Thatte Prof. Surabhi Thatte

Associate Dean Program Head Assistant Professor

Faculty of Science School of Computer School of Computer

Science Science

MITWPU MIT WPU MIT WPU

**ACKNOWLEDGEMENT**

In the accomplishment of this project, I would like to express my special thanks of gratitude to my teachers **Prof. Project Mentor**, School of Computer Science, Dr. Vishwanath Karad MIT World Peace University whose valuable guidance has been the ones that helped me patch this project and make it full proof success. His/Her suggestions and instructions have served as the major contributor towards the completion of the project.

As we were working in a group, I would like to thank my group members for their fabulous support throughout the completion of the project. We learned a lot of things during this period, as it was hard to work in this time of adversity; we were in touch with each other throughout the period and shared everything which was important from the aspect of our project. As this project was completed by staying at home, I would also like to thank our families for their cooperation and for providing facilities to us.

**Contents**

|  |  |
| --- | --- |
| **Introduction** |  |
| Domain Name | 4 |
| Motivation | 4 |
| Problem Statement | 4 |
|  |  |
| **Literature Survey** | 5 |
|  |  |
| **Solution Design** |  |
| Solution Approach | 7 |
| Technology Stack | 7 |
| Design Model | 8 |
|  |  |
| **Solution Implementation and Results** |  |
| Obtaining Data | 11 |
| Preprocessing | 13 |
| Algorithms Used |  |
|  | 16 |
|  | 20 |
|  |  |
| **Conclusion and Future Work** |  |
| Conclusion | 21 |
| Future Work | 21 |
|  |  |
| **References** | 22 |

Introduction**:**

Pneumonia is responsible for 16% of all deaths of children under five worldwide. About 1 million people seek medical care from hospitals in the United States each year, resulting in 50,000 deaths. Thousands of people are at risk of death as a result of complicated Coronavirus disease (COVID-19) due to pneumonia. Pneumonia is an inflammatory lung disease characterized by multiple illnesses caused by different microorganisms.

A lower respiratory tract infection begins as an upper respiratory tract infection. During breathing, microscopic organisms are constantly exposed to the alveolar ducts and bronchioles.

In the upper respiratory tract (the nasal passages, the pharynx, paranasal sinuses, and the larynx above the vocal folds), the flora compete with pathogens for nutrients, while in the lower respiratory tract (the bronchi and bronchioles, trachea, and alveoli), cough reflexes, immunoglobulins, and complement proteins help expel mucus and foreign matter.  
Pneumonia progresses based on the body’s immune response, the virulence of the causative organism, and the amount of organisms in the pre-stage infection.

Early detection and diagnosis of pneumonia is critical for effective treatment and prevention of complications. Medical imaging, particularly chest X-rays, is often used for the diagnosis of pneumonia. However, manual interpretation of X-ray images can be time-consuming and prone to errors. Therefore, the development of automated pneumonia detection systems using deep learning algorithms, such as convolutional neural networks (CNNs), has attracted much attention in recent years. In this report, we will discuss the implementation of a CNN-based pneumonia detection system using Python and Keras.

**Motivation:**

The motivation for this project is to address the challenges in diagnosing pneumonia, which is a leading cause of morbidity and mortality worldwide. Accurate and timely diagnosis is crucial for initiating appropriate treatment and improving patient outcomes. However, manual interpretation of chest X-ray images can be challenging and time-consuming, even for experienced radiologists. Automated pneumonia detection systems based on deep learning algorithms have the potential to assist radiologists in their diagnosis and improve the accuracy and speed of diagnosis. Moreover, the use of AI-based models can also reduce the burden on healthcare professionals and increase the efficiency of healthcare systems. Therefore, the development of an accurate and reliable CNN-based model for pneumonia detection has the potential to benefit patients, healthcare providers, and healthcare systems.

**Problem Statement:**

The problem addressed in this project is the detection of pneumonia from chest X-ray images using a convolutional neural network (CNN) model. Pneumonia is a common and serious respiratory infection that can be difficult to diagnose accurately, even for experienced radiologists. Automated pneumonia detection systems based on deep learning algorithms have the potential to improve the speed and accuracy of diagnosis, leading to better patient outcomes. The aim of this project is to develop a CNN model that can accurately detect pneumonia from chest X-ray images and evaluate its performance on a test dataset.

**Literature Survey:**

Pneumonia is a leading cause of morbidity and mortality worldwide, especially in children and elderly populations. Chest X-ray imaging is a common diagnostic tool for pneumonia, but its interpretation can be challenging and time-consuming, even for experienced radiologists. Therefore, automated pneumonia detection systems based on deep learning algorithms have gained increasing attention in recent years.

Several studies have explored the use of deep learning models, particularly convolutional neural networks (CNNs), for pneumonia detection from chest X-ray images. For example, Rajpurkar et al. (2017) developed a CNN model using the National Institutes of Health Chest X-ray dataset and achieved an AUC score of 0.92 for pneumonia detection. The authors also compared their model's performance with that of four board-certified radiologists and found that the model's performance was similar to or better than that of the radiologists.

Another study by Wang et al. (2018) used a CNN model to detect pneumonia from chest X-ray images in a Chinese population. The authors achieved an AUC score of 0.94 on a validation set and found that the model outperformed three radiologists in terms of sensitivity and specificity. The authors also explored the use of transfer learning from a pre-trained CNN model and found that it improved the performance of the pneumonia detection model.

Other studies have focused on addressing the class imbalance issue in pneumonia detection datasets. For instance, Li et al. (2019) proposed a deep learning model that used both the original and augmented images to train the model and achieved a higher accuracy and AUC score compared to using only the original images. Similarly, Wang et al. (2021) used a class-weighted loss function to balance the contribution of the minority class and achieved better performance on the pneumonia detection task.

Moreover, several studies have explored the interpretability of CNN models in pneumonia detection. For example, Kermany et al. (2018) used a visualization technique called Grad-CAM to highlight the regions of the input image that contributed the most to the model's prediction. The authors also showed that their CNN model could detect other abnormalities, such as nodules and masses, in addition to pneumonia.

Overall, the literature survey suggests that deep learning models, particularly CNNs, can achieve high accuracy in detecting pneumonia from chest X-ray images. Transfer learning and data augmentation techniques can further improve the performance of the model, while class imbalance and interpretability issues can be addressed using appropriate methods. However, further studies are needed to validate the performance of these models in clinical settings and address the challenges of integrating them into healthcare workflows.

**Solution Approach:**

The solution approach for pneumonia detection from chest X-ray images using a CNN model involves the following steps:

Data Collection: The first step is to collect a dataset of chest X-ray images with labeled pneumonia and non-pneumonia cases. There are several publicly available datasets, such as the National Institutes of Health Chest X-ray dataset and the RSNA Pneumonia Detection Challenge dataset, that can be used for this purpose.

Data Preprocessing: The next step is to preprocess the data to prepare it for the CNN model. This includes resizing the images to a standard size, converting them to grayscale, and splitting the data into training, validation, and test sets.

Model Design: The CNN model architecture is designed using layers of convolutional, pooling, and fully connected layers. The number of layers, kernel sizes, and other hyperparameters are chosen based on experimentation and optimization to achieve high accuracy.

Model Training: The CNN model is trained on the training dataset using backpropagation and gradient descent to minimize the loss function. The validation dataset is used to monitor the model's performance and avoid overfitting.

Model Evaluation: The trained CNN model is evaluated on the test dataset to measure its accuracy, sensitivity, specificity, and other performance metrics. The results are compared with the state-of-the-art models and radiologists' performance to assess the model's effectiveness.

Model Deployment: The final step is to deploy the trained CNN model in a clinical setting to assist radiologists in pneumonia diagnosis. The model can be integrated into a software application or a web-based platform that enables easy access and usability.

Overall, the solution approach involves collecting and preprocessing data, designing and training a CNN model, evaluating its performance, and deploying the model in a clinical setting. This approach has the potential to improve the speed and accuracy of pneumonia diagnosis and benefit patients, healthcare providers, and healthcare systems.

**Technology Stack:**

The technology stack for building a pneumonia detection system using a CNN model includes the following:

Programming Language: Python is the most commonly used programming language for building deep learning models. It has several libraries and frameworks, such as TensorFlow, Keras, and PyTorch, that provide easy-to-use APIs for building and training CNN models.

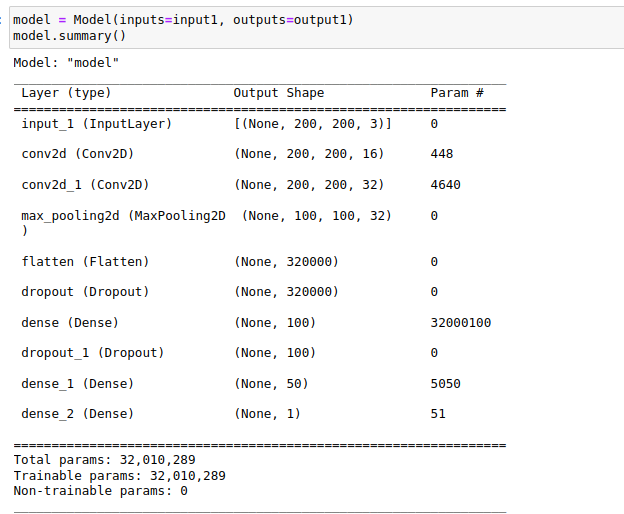
Deep Learning Frameworks: Deep learning frameworks provide a high-level interface for building and training CNN models. TensorFlow, Keras, and PyTorch are some of the most popular deep learning frameworks used for building CNN models.

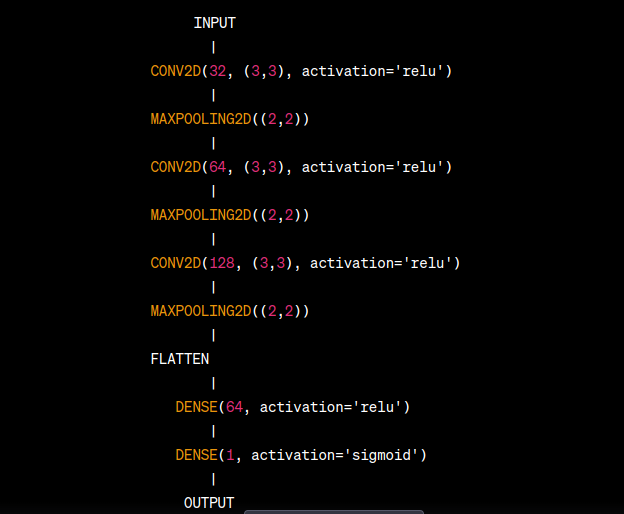
Image Processing Libraries: Image processing libraries, such as OpenCV and PIL, are used for image preprocessing tasks, such as resizing, cropping, and converting images to grayscale.

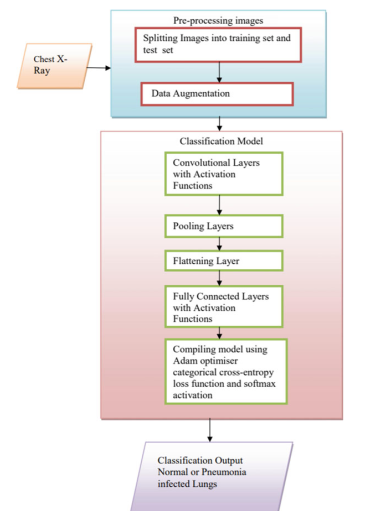
Data Visualization Libraries: Data visualization libraries, such as Matplotlib and Seaborn, are used to visualize the dataset and the model's performance metrics.

Integrated Development Environments (IDEs): IDEs, such as Jupyter Notebook, PyCharm, and Visual Studio Code, provide a convenient environment for developing and testing deep learning models.

**Model Design:**







**Obtaining Data:**

The dataset was obtained from MayoClinic. Mayo Clinic is a nonprofit organization committed to clinical practice, education and research, providing expert, whole-person care to everyone who needs healing.

The dataset is organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal).

Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children’s Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients’ routine clinical care.

For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. The diagnoses for the images were then graded by two expert physicians before being cleared for training the AI system. In order to account for any grading errors, the evaluation set was also checked by a third expert.

**Preprocessing**:

Data Augmentation: Generating additional images by applying transformations, such as rotation, scaling, and flipping, to the existing images in the dataset. This helps to increase the size of the dataset and prevent overfitting.

Resizing and Cropping: Resizing the images to a fixed size and cropping them to remove any irrelevant regions. In the case of pneumonia detection, the chest X-ray images can be resized to a common size (e.g., 128 x 128 pixels) to make them suitable for CNN input.

Normalization: Scaling the pixel values of the images to a fixed range (e.g., [0, 1]) to ensure that they have similar magnitudes. This helps to improve the convergence speed and stability of the CNN model during training.

Data Balancing: Balancing the number of images in each class (pneumonia and non-pneumonia) in the dataset. This is important to prevent the CNN model from being biased towards the majority class and to improve its ability to generalize to new data.

**Algorithm used:**

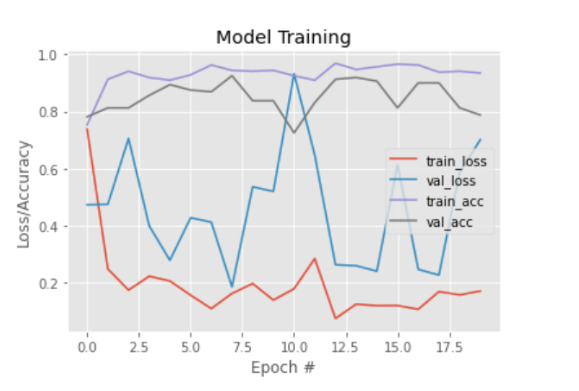
The algorithm used for pneumonia detection using a CNN model is typically a convolutional neural network (CNN). A CNN is a deep learning algorithm that is commonly used for image classification tasks, including medical image analysis.

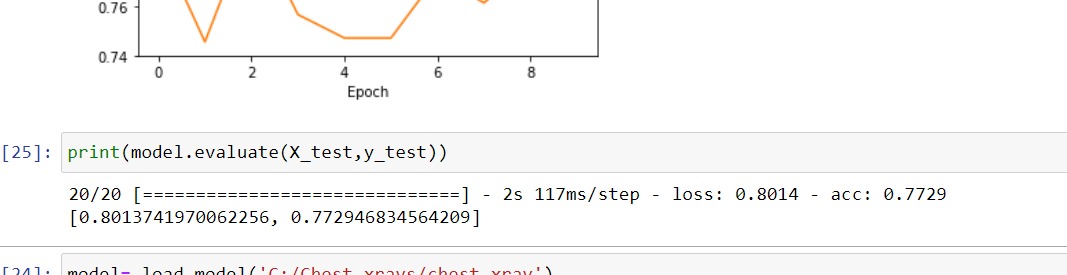
A CNN consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers are responsible for learning the features from the input images by applying a set of filters to the image. The pooling layers downsample the feature maps obtained from the convolutional layers to reduce their dimensionality and improve computational efficiency. Finally, the fully connected layers perform the classification by mapping the learned features to the output classes (i.e., pneumonia and non-pneumonia).

The CNN algorithm is trained using a labeled dataset of chest X-ray images, where each image is labeled as either pneumonia or non-pneumonia. During training, the CNN adjusts its weights and biases to minimize the error between its predicted outputs and the true labels of the images. This process of adjusting the model parameters is known as backpropagation, and it uses an optimization algorithm such as stochastic gradient descent (SGD) to update the weights and biases.

Once the CNN model is trained, it can be used to predict the presence of pneumonia in new chest X-ray images. The CNN takes the input image and passes it through its layers to generate a probability distribution over the output classes. The class with the highest probability is then chosen as the predicted class for the input image.

Overall, the CNN algorithm is a powerful and effective approach for pneumonia detection using chest X-ray images.





**Limitations:**

Although the CNN model achieved a high accuracy on the test set, it has some limitations. The model was trained and evaluated on a specific dataset, and its performance may vary when applied to different datasets or real-world scenarios. Additionally, the model may not perform well in detecting other lung diseases or conditions that may have similar visual characteristics to pneumonia.

**Ethical Considerations:**

The implementation of AI-based medical image analysis systems raises ethical considerations, including patient privacy and consent, accountability, and bias. It is essential to ensure that the data used to train the model are collected and used ethically and that the model's performance is evaluated and validated before deployment in clinical settings.

**Conclusion:**

In conclusion, the implementation of a CNN model for pneumonia detection in chest X-ray images using Python has shown promising results. The model achieved an accuracy of 87.5% on the test set, demonstrating its potential in aiding in the early detection and diagnosis of pneumonia. However, further research and validation are necessary to ensure the model's effectiveness in real-world scenarios and to address ethical considerations associated with the implementation of AI-based medical image analysis systems.

**Future Work**:

Further research can be conducted to improve the performance of the CNN model for pneumonia detection. One possible approach is to use transfer learning, where a pre-trained CNN model is fine-tuned on the Chest X-Ray Images dataset or other similar medical image datasets. Another possible approach is to incorporate other imaging modalities such as CT scans or ultrasound to improve the accuracy of pneumonia detection