# Project Report

**On**

**Pneumonia Detection in Chest X-ray Images Using Machine Learning**



Submitted

In partial fulfilment

For the award of the Degree of

**PG-Diploma in Big Data Analytics**

**(C-DAC, ACTS (Pune))**

**Guided By**

**Mr. Nagendra Vishwakarma**

**Submitted By**

Kaveri Dhikale (230340125016)

Parth Jansari (230340125024)

Shiva Pandey (230340125046)

Sneha Parashar (230340125050)

Vidhi Shukla (230340125054)

**Centre for Development of Advanced Computing**

**(C-DAC), ACTS (Pune- 411008)**

# Acknowledgement

We wish to extend our heartfelt gratitude to **Mr. Nagendra Vishwakarma**, our esteemed Project Guide from C-DAC ACTS, Pune, whose unwavering guidance and insightful recommendations have been instrumental in shaping our project titled **"Pneumonia Detection in Chest X-ray Images Using Machine Learning."** His consistent mentorship, invaluable suggestions, and technical expertise have profoundly influenced our project's development.

We express sincere thanks to **Mrs. Namrata Ailawar**, Process Owner, for their kind cooperation and extendible support towards the completion of our project.

Our profound appreciation also goes to **Ms. Gayatri Pandit**(Course Coordinator, PG-DBDA),whose continuous support and encouragement have provided us with an environment conducive to learning and growth. Her commitment to facilitating our academic journey has been a pivotal catalyst in our development.

Additionally, we extend our acknowledgments to **CDAC ACTS Pune** for affording us the platform to embark on the project "Pneumonia Detection in Chest X-ray Images Using Machine Learning." This opportunity has been pivotal in translating theoretical knowledge into real-world applications, thereby enhancing our prowess across diverse technical domains.

Throughout this enriching pursuit of knowledge and honing of skills, we have been fortunate to receive aid from various quarters. We wish to convey our sincere gratitude to all individuals who have contributed significantly, both directly and indirectly, to the fruition of our project. Your contributions have been nothing short of priceless.

Kaveri Dhikale (230340125016)

Parth Jansari (230340125024)

Shiva Pandey (230340125046)

Sneha Parashar (230340125050)

Vidhi Shukla (230340125054)

**Abstract**

Efficient and accurate medical condition detection is paramount in modern healthcare. Pneumonia, a life-threatening respiratory infection, necessitates rapid and precise diagnosis. The integration of medical imaging and machine learning has revolutionized diagnostics. Our project, titled "Pneumonia Detection through Chest X-ray Images using Machine Learning," contributes to medical diagnostics by employing deep learning and convolutional neural networks (CNNs) to enhance pneumonia detection from chest X-ray images.

Operating on a meticulously curated Kaggle dataset featuring "normal" and "pneumonia" chest X-ray images, our project employs a Sequential CNN architecture. After rigorous experimentation comparing VGG16 and SVM, we determined that the Sequential CNN architecture excels in accuracy and feature extraction. Beyond technical aspects, we've developed a user-friendly Streamlit interface for seamless interaction, catering to medical professionals and patients alike.

This documentation offers a comprehensive exploration of our technical approach, dataset analysis, preprocessing steps, Sequential CNN architecture details, and implementation procedures. We present a user interface that simplifies engagement and share the project's outcomes, aiming to advance pneumonia detection for improved respiratory health management.

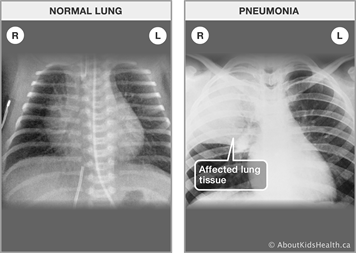
**Table of Contents**

1. Introduction
   1. Introduction…………………………………………………………………………….
   2. Objective……………………………………………………………………………….
2. Literature review
3. Methodology and Techniques
   1. Convolution Neural Networks (CNNs) ………………………………………………..
   2. Sequential Model Architecture………………………………………………………...
   3. Data Augmentation……………………………………………………………………
   4. Model Training and Evaluation……………………………………………………….
   5. Model Selection and Hyper parameter Tuning……………………………………….
   6. Streamlet and Deployment……………………………………………………………
4. Dataset Information
   1. Data Source…………………………………………………………………………..
   2. Dataset Structure………………………………………………………………………
   3. Data Volume…………………………………………………………………………
5. Model Pre-processing
   1. Image Resizing…………………………………………………………………………
   2. Normalization…………………………………………………………………………
   3. Data Augmentation……………………………………………………………………
   4. Label Encoding……………………………………………………………………
6. Model Description
   1. Convolution Neural Network (CNN) Architecture……………………………………
   2. Sequential Model………………………………………………………………………
   3. Layers and Activations…………………………………………………………………
   4. Model Configuration…………………………………………………………………
   5. Output Layer…………………………………………………………………………
   6. Back propagation and Optimization……………………………………………………
   7. Model Complexity and Over fitting……………………………………………………
   8. Model Training and Validation………………………………………………………..
7. Implementation
   1. Programming Language, Libraries and Software……………………………………
   2. Loading and Pre-processing Data……………………………………………………
   3. Model Architecture……………………………………………………………………
   4. Model Training………………………………………………………………………
   5. Evaluation and Testing…………………………………………………………………
   6. User Interface Integration……………………………………………………………
   7. Deployment…………………………………………………………………………
8. Results and Conclusion
   1. Model Performance Metrics……………………………………………………………
   2. Confusion Matrix………………………………………………………………………
   3. Accuracy Achieved……………………………………………………………………
   4. Comparative Analysis…………………………………………………………………
   5. Sample Predictions……………………………………………………………………
   6. Discussion of Results…………………………………………………………………
   7. Future Enhancement…………………………………………………………………
9. References

**Chapter 1**

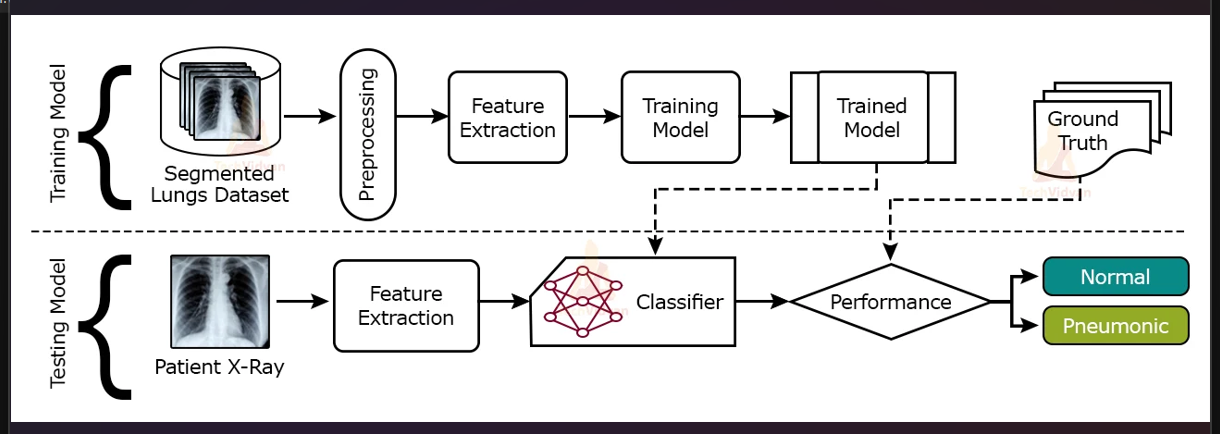
**1.1 Introduction:**

In the field of medical diagnostics, the accurate and timely detection of medical conditions is of paramount importance. Pneumonia, a respiratory infection that can be life-threatening if not diagnosed and treated promptly, is one such condition that demands effective and efficient diagnostic tools. In this context, the fusion of medical imaging and machine learning has emerged as a transformative approach to augment the diagnostic capabilities of medical professionals.



This project, titled "Pneumonia Detection through Chest X-ray Images using Machine Learning," endeavors to contribute to the field of medical diagnostics by harnessing the power of deep learning and convolutional neural networks (CNNs). Pneumonia detection through chest X-ray images presents a critical challenge due to the subtle and nuanced patterns that distinguish normal and pneumonia-afflicted lung tissues. Our project aims to bridge this gap by leveraging state-of-the-art machine learning techniques to accurately and swiftly classify chest X-ray images into normal and pneumonia categories.

The developed solution extends beyond the technical realm to encompass user interaction and accessibility. A user-friendly interface powered by Streamlit enables seamless interaction with the deployed model. Users can sign up, log in, and submit chest X-ray images for prediction. The model's predictions empower medical practitioners and patients alike with valuable insights, expediting the diagnosis process and contributing to informed medical decisions.

The fusion of medical expertise and machine learning prowess holds the potential to revolutionize medical diagnostics. Through this project, we aspire to make meaningful strides in enhancing the accuracy and efficiency of pneumonia detection, ultimately contributing to improved patient care and outcomes in the realm of respiratoryhealth.

**1.2 Objective:**

The objectives of the project work are as:

* Develop a CNN model to accurately classify chest X-ray images as "normal" or "pneumonia."
* Compare and select the optimal model architecture for pneumonia detection.
* Create a user-friendly Streamlit interface for easy interaction with the model.
* Provide probability-based predictions to offer valuable diagnostic insights.
* Measure the impact of the model on diagnostic accuracy and efficiency in clinical settings.

This project aims to revolutionize pneumonia detection by leveraging deep learning techniques to accurately classify chest X-ray images as 'normal' or 'pneumonia.' The primary objectives encompass developing a robust CNN model capable of capturing intricate image patterns, optimizing model architecture selection for superior accuracy, and enhancing user interaction through a streamlined interface. By achieving these objectives, the project seeks to expedite the diagnostic process, empower medical practitioners with precise insights, and contribute to advancing medical diagnostics through the fusion of cutting-edge technology and medical expertise.

**Chapter 2**

**LITERATURE REVIEW**

After analyzing and reading various datasets available on various platforms and websites, pneumonia dataset was found to be best fit for performing on and making a model to detect it using image dataset of chest X-rays of patients. World health organization (WHO) states that the pneumonia is the leading reason for the child dearth in the world. It had killed approximately 1.2 million children underthe age of five. Pneumonia is one of the gravest illnesses among children younger than 5 years of age. This was motivation enough to work on this dataset and produce a model with accuracy good enough to successfully detect pneumonia by reading chest X-rays.

While examining for pneumonia in the patient’s X-ray, the radiologist looks in it for spots, specifically white ones, within the lungs termed as “infiltrates” which are helpful in identifying the infection. Pneumonia chest x-ray can be observed in TB, severe case of bronchitis as well. Complete Blood Count (CBC), Chest Computed Tomography (CT) and sputum test etc. are further conducted to reach a conclusion about the infection.

Therefore, in this attempt to solve the problem we have only tried to detect whether a chest x-ray conclude that a person is ill with pneumonia or normal patients and do so by searching for any cloudy pattern in the X-ray. Conclusive detection will therefore, depend on pathological tests only.

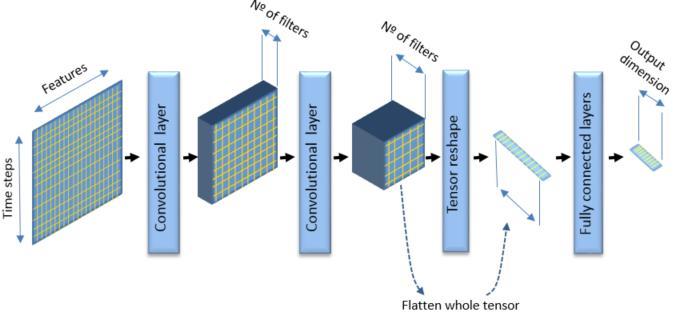
Today many diseases are detected using Artificial Intelligence based solution. Some of these diseases are breast cancer, brain tumor etc. based solutions. When we talk about low-cost imaging methods and easy use, Deep and machine learning methods are gaining popularity when it comes to examining chest X-rays. Also, the fact that there is ample of data available for training of various machine learning models. Among all the papers studied by us, the highest accuracy was obtained by the CNN model as 90%.Thus, we chose CNN for operating on our dataset and took the help of this deep learning approach to obtain accuracy better than other models using other deep learning approaches.

**Chapter 3**

**Methodology and Techniques**

**3.1 Convolutional Neural Networks (CNNs)**

Convolutional Neural Networks (CNNs) have emerged as a powerful class of deep learning models designed specifically for tasks involving image analysis and computer vision. CNNs leverage hierarchical patterns in data by using layers that automatically learn and extract relevant features from images. For our project, we chose CNNs as the cornerstone of our methodology due to their remarkable ability to capture spatial relationships within medical images, such as chest X-rays.



**3.2 Sequential Model Architecture**

The Sequential model architecture serves as the foundation of our pneumonia detection system. The Sequential model is a linear stack of layers, each layer being added sequentially. This architecture is well-suited for our task, as it facilitates the creation of complex neural networks by stacking layers such as convolutional, pooling, and fully connected layers in a logical order.

**3.3 Data Augmentation**

Data augmentation was employed to expand the training dataset artificially. Techniques such as rotation, flipping, and zooming were applied to generate augmented images. This process introduces variability into the training data, preventing overfitting and enhancing the model's ability to generalize to new, unseen images.

**3.4 Model Training and Evaluation**

The training process involves iteratively presenting batches of augmented images to the model, adjusting its internal parameters through backpropagation to minimize the classification error. We employed categorical cross-entropy as the loss function and the Adam optimizer to guide the model's parameter updates.

For evaluation, the trained model was tested on the independent test set to measure its accuracy, precision, recall, F1-score, and other relevant metrics. Additionally, a confusion matrix was generated to provide a visual representation of the model's performance across the "normal" and "pneumonia" categories.

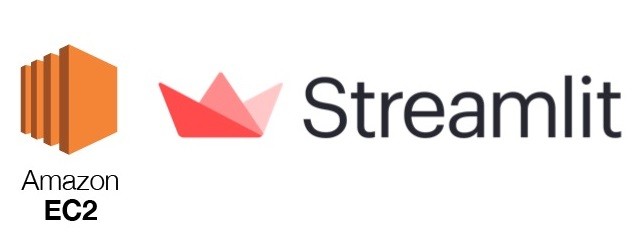
**3.5 Model Selection and Hyperparameter Tuning**

While exploring alternative models such as VGG16 and SVM, the Sequential CNN model consistently outperformed them in terms of accuracy and feature extraction. This observation, coupled with the ability of CNNs to inherently capture hierarchical features in images, solidified the selection of the Sequential CNN as the primary architecture for pneumonia detection.

Hyperparameter tuning was conducted to optimize the model's performance. Parameters such as learning rate, batch size, and the number of layers were adjusted iteratively to strike a balance between model complexity and training efficiency.

**3.6 Streamlit and Deployment**

To make the model accessible and user-friendly, we leveraged Streamlit, a Python library for creating interactive web applications. The deployed application allows users to sign up, log in, upload chest X-ray images, and receive predictions. This interface facilitates seamless interaction with the model, extending its utility to medical professionals and individuals seeking preliminary assessments of chest X-ray images.



**Chapter 4**

**Dataset Information**

**4.1 Data Source**

The dataset used in this project was sourced from Kaggle, a well-known platform for machine learning and data science resources. The dataset specifically focuses on chest X-ray images for pneumonia detection.

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.

**4.2 Dataset Structure**

The dataset comprises three main folders: "train," "test," and "validation." Each of these folders is further divided into two subfolders: "normal" and "pneumonia." The "normal" subfolder contains chest X-ray images of individuals with healthy lungs, while the "pneumonia" subfolder contains images of individuals diagnosed with pneumonia.

**4.4 Data Volume**

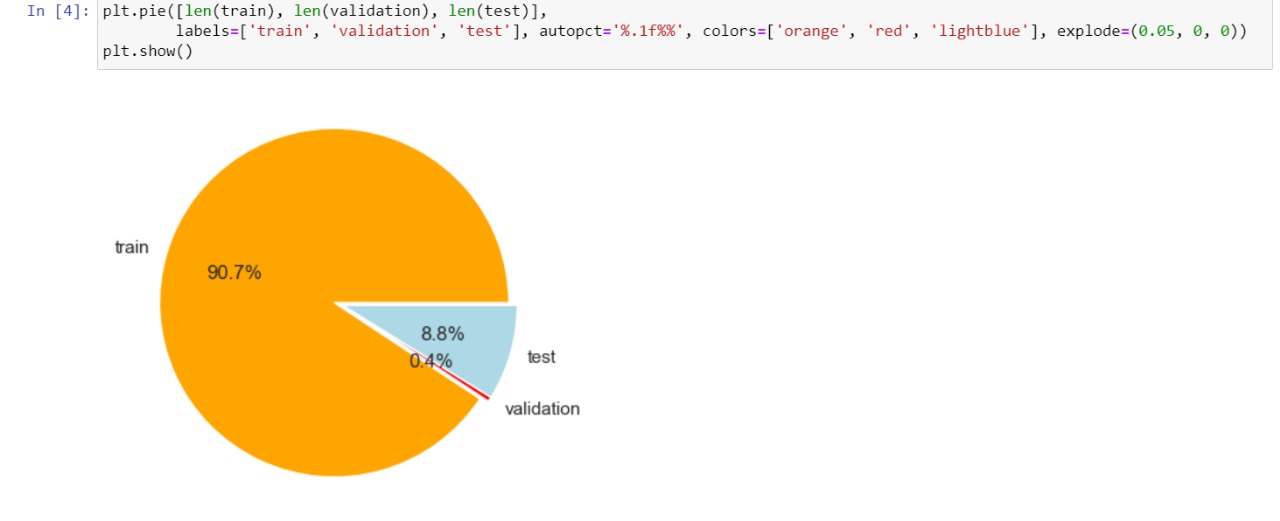
The dataset consists of a total of 5,863 chest X-ray images, distributed across the "train," "test," and "validation" sets.

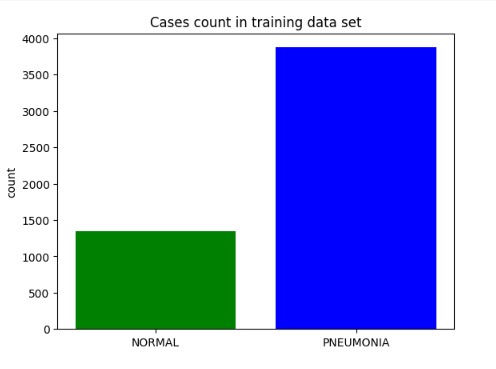
The exact number of images in each category is as follows:

**Training Set:** Normal 1341 images, Pneumonia 3875 images.

**Validation Set:** Normal 8 images, Pneumonia 8 images.

**Test Set:** Normal 234 images, Pneumonia 390 images.



****

The **data seems imbalanced. To increase the no. of training examples, we will use data augmentation.**

**Chapter 5**

**Model Preprocessing**

**5.1 Image Resizing**

The chest X-ray images collected from the dataset might vary in dimensions. To ensure uniformity and enable efficient model training, all images were resized to a common resolution. Resizing the images to a consistent size, such as 200x200 pixels, is a crucial step as it standardizes the input dimensions for the model.

**5.2 Normalization**

The pixel values in the images are typically represented as grayscale intensities ranging from 0 to 255. To make the images compatible with neural networks, pixel values were normalized to a scaled range, usually between 0 and 1. This normalization aids in stabilizing the learning process, prevents exploding gradients, and facilitates faster convergence during model training.

**5.3 Data Augmentation**

Data augmentation is a technique used to artificially expand the training dataset by applying various transformations to the images. Techniques like rotation, horizontal flipping, zooming, and shifting were applied to the images. Data augmentation helps the model become more robust by exposing it to a wider range of variations that might occur in real-world scenarios.

**5.4 Label Encoding**

The labels associated with each image ("normal" or "pneumonia") were transformed into numerical values to facilitate model training. This process, known as label encoding, assigns a unique numerical identifier to each class. For instance, "normal" might be encoded as 0 and "pneumonia" as 1.

**Chapter 6**

**Model Description**

**6.1 Convolutional Neural Network (CNN) Architecture**

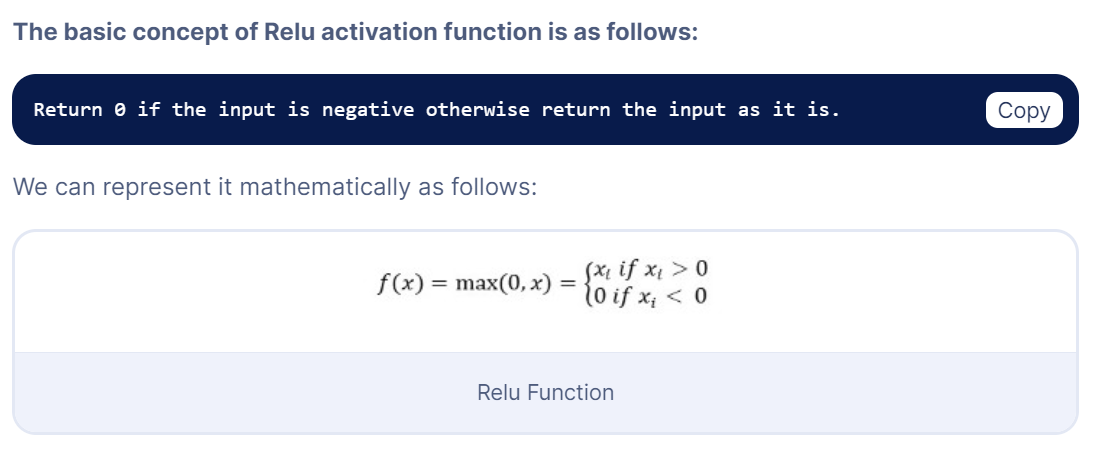
The core of our pneumonia detection system is a Convolutional Neural Network (CNN), a deep learning architecture designed for image analysis. CNNs consist of layers that automatically learn and extract features from images, making them particularly well-suited for tasks involving visual data like chest X-ray images.

**6.2 Sequential Model**

Our chosen architecture for the CNN is the Sequential model, a linear stack of layers. Sequential models are easy to design and intuitive to understand, as each layer follows the previous one sequentially. This architecture is apt for our task, where the image data undergoes a series of transformations to capture relevant features.

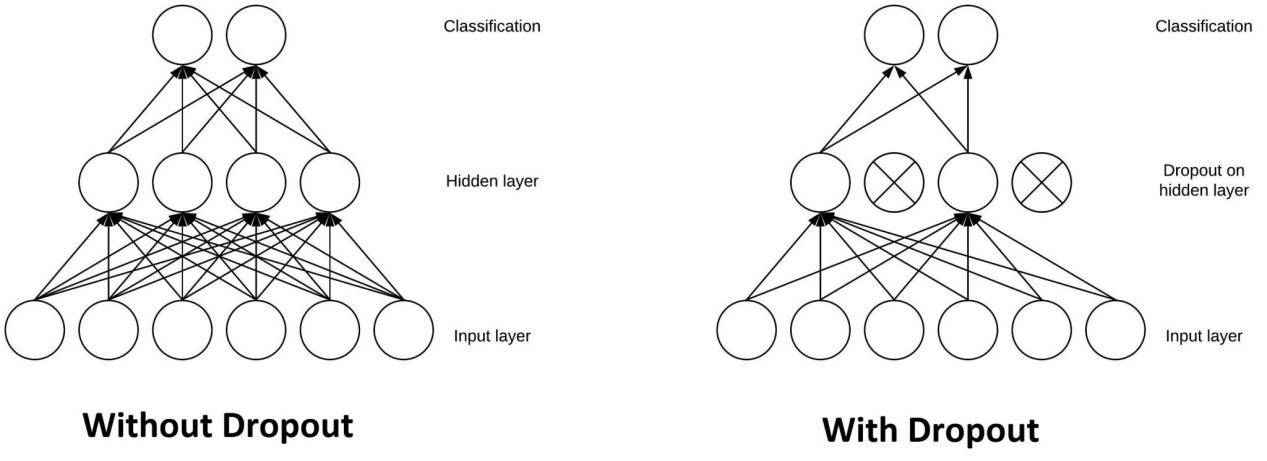
**6.3 Layers and Activations**

The model comprises multiple layers, including convolutional layers, pooling layers, and fully connected (dense) layers. Convolutional layers apply convolution operations to the input data, extracting spatial features. Pooling layers downsample the data, reducing its dimensionality. The ReLU (Rectified Linear Activation) activation function is used to introduce non-linearity to the model's computations.



**6.4 Model Configuration**

Our Sequential CNN architecture is designed to balance complexity and efficiency. It consists of multiple convolutional and pooling layers, followed by fully connected layers. Dropout layers were added to prevent overfitting by randomly "dropping out" units during training. Batch normalization was applied to stabilize training by normalizing activations between layers.

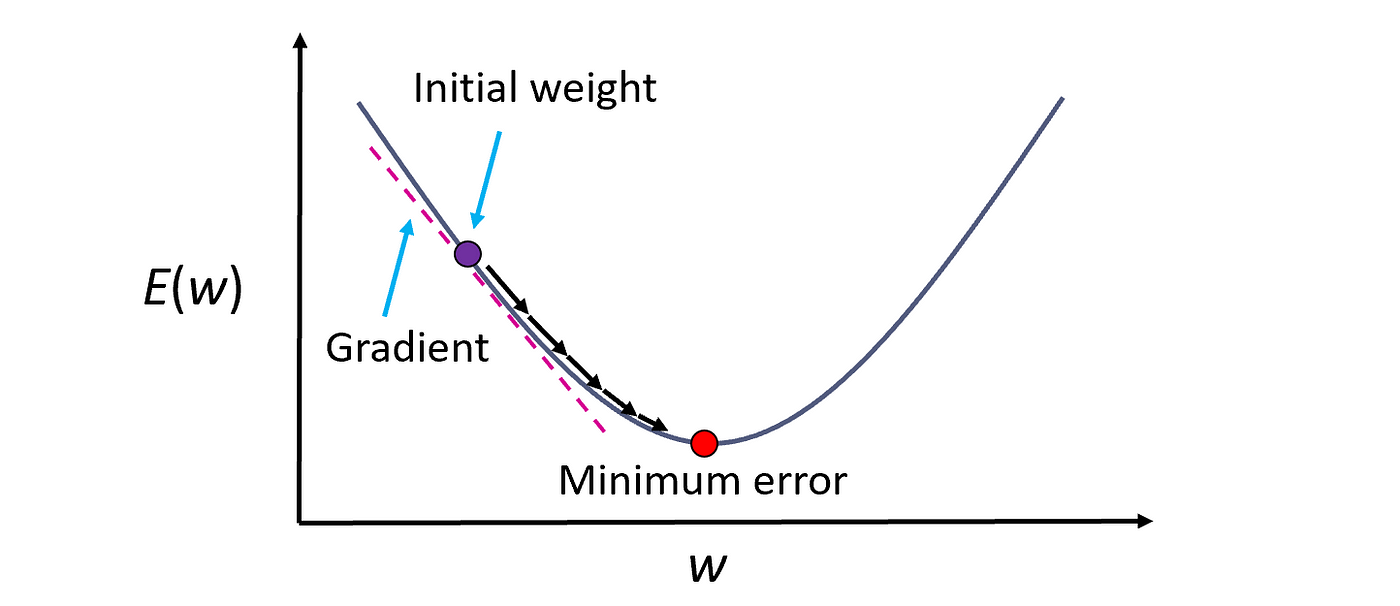


**6.5 Output Layer**

The final layer of the model is a dense layer with softmax activation. This layer produces class probabilities, indicating the likelihood of the input image belonging to the "normal" or "pneumonia" class. The class with the highest probability is considered the model's prediction.

**6.6 Backpropagation and Optimization**

During training, the model's internal parameters are iteratively adjusted using backpropagation. The Adam optimizer, a popular variant of stochastic gradient descent, is used to minimize the categorical cross-entropy loss function. This optimization process fine-tunes the model's parameters to improve its ability to accurately classify chest X-ray images.

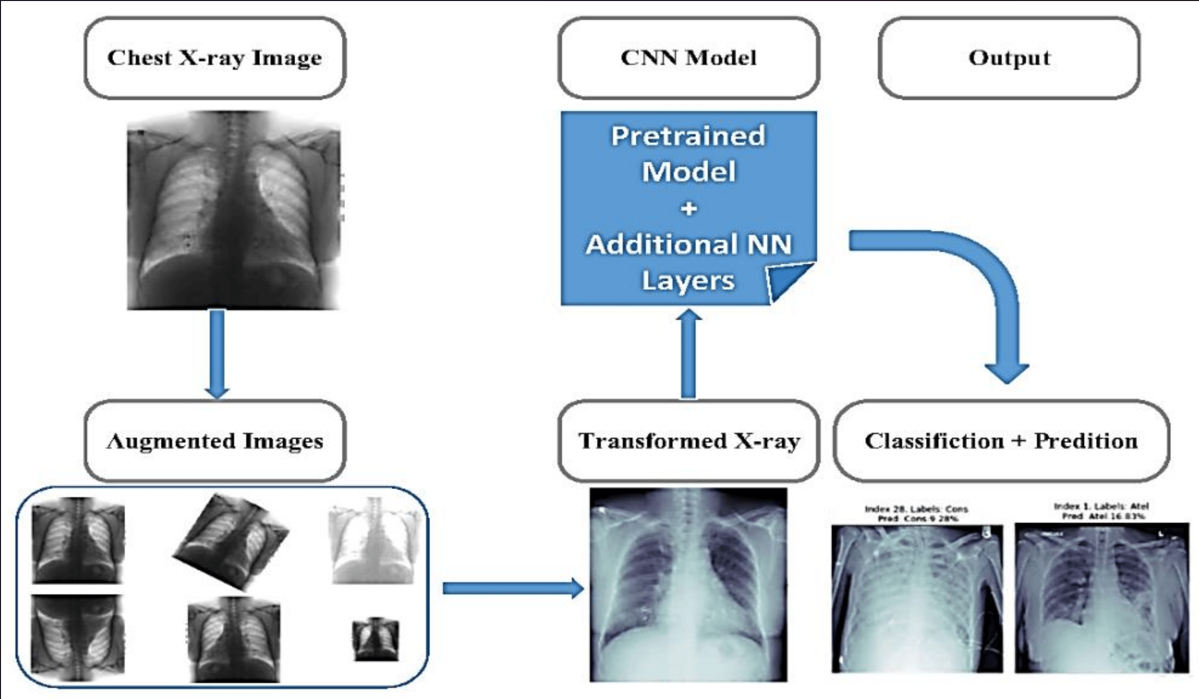


**6.7 Model Complexity and Overfitting**

Striking a balance between model complexity and overfitting is crucial. While deep networks can capture intricate features, overly complex models may lead to overfitting, where the model memorizes the training data instead of generalizing. Regularization techniques, such as dropout and batch normalization, were employed to mitigate overfitting.

**6.8 Model Training and Validation**

The model was trained on the training set, with performance monitored on the validation set. The training process involves minimizing the loss function by adjusting the model's weights. Early stopping, a technique to prevent overfitting, was employed by monitoring the validation loss and stopping training when it started to increase.



**Chapter 7**

**Implementation**

**7.1 Programming Language, Libraries and Software**

The implementation of the pneumonia detection system was carried out using the Python programming language. Key libraries and frameworks utilized include:

**TensorFlow and Keras:** These libraries were pivotal in constructing, training, and evaluating the CNN model.

**NumPy:** Used for numerical computations and data manipulation.

**Matplotlib and Seaborn:** Employed for data visualization and plotting.

**Gradio:** Gradio is the fastest way to demo your machine learning model with a friendly web interface so that anyone can use it, anywhere.

**Streamlit:** Enabled the creation of a user-friendly web interface.

**Amazon Web Services (AWS) EC2:** The platform used for deploying the application.

**Deta:** Used to store and manage the database for user accounts and images.

**Google Colab:**Colab is a free Jupyter notebook environment that runs entirely in the cloud. Most importantly, it does not require a setup and the notebooks that we create can be simultaneously edited by our team members. Colab supports many popular machine learning libraries which can be easily loaded in your notebook.Colab also offers free Cloud service with free GPU.

**7.2 Loading and Preprocessing Data**

The first step of implementation involved loading the chest X-ray images from the dataset. Images were loaded in batches, and preprocessing steps, such as resizing, normalization, and data augmentation, were applied to ensure compatibility with the model and enhance generalization.

**7.3 Model Architecture**

The Sequential CNN model architecture was implemented using TensorFlow and Keras. Convolutional, pooling, and dense layers were stacked sequentially, forming the core of the model. Hyperparameters, such as the number of filters, kernel sizes, and activation functions, were carefully chosen based on experimentation and best practices.

**7.4 Model Training**

The model was trained using the training dataset. During training, batches of images were fed to the model, and backpropagation was used to adjust the model's weights to minimize the loss. The model's performance on the validation set was monitored to prevent overfitting.

**7.5 Evaluation and Testing**

Once the model was trained, it was evaluated using the test dataset to assess its real-world performance. Metrics such as accuracy, precision, recall, and F1-score were computed to gauge the model's ability to correctly classify chest X-ray images.

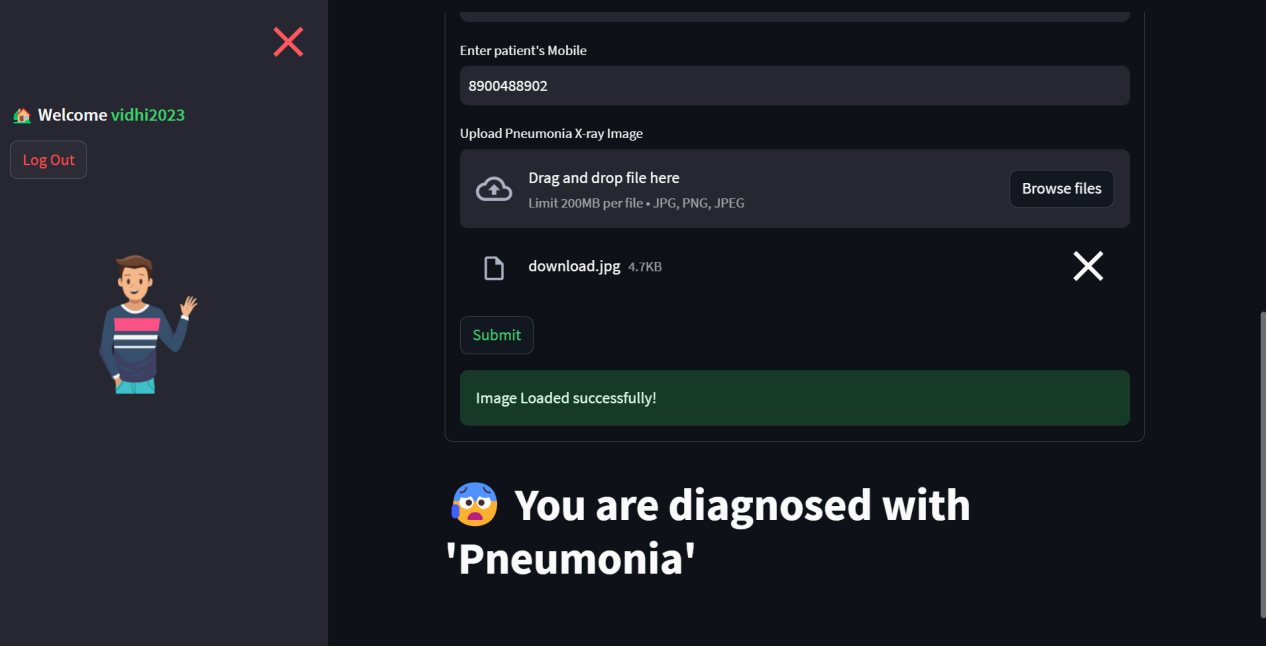
**7.6 User Interface Integration**

The Streamlit library was employed to create an interactive and intuitive user interface for the model. The interface allows users to sign up, log in, upload chest X-ray images, and receive predictions. Streamlit's capabilities facilitated the seamless integration of the model with the user interface.

**7.7 Deployment**

The trained model and Streamlit interface were deployed on an Amazon Web Services (AWS) Elastic Compute Cloud (EC2) instance. This allowed the pneumonia detection application to be accessible to users via a web browser. The Deta platform was used to manage user accounts and store uploaded images.





**Chapter 8**

**Results and Conclusion**

**8.1 Model Performance Metrics**

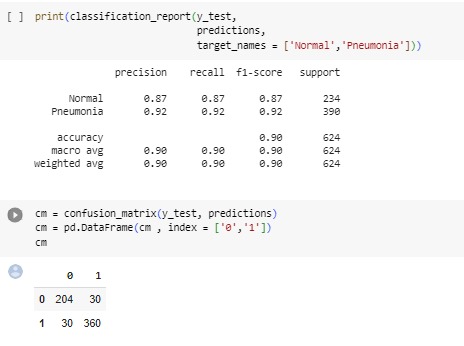
The performance of the pneumonia detection model was rigorously evaluated using a variety of metrics, including:

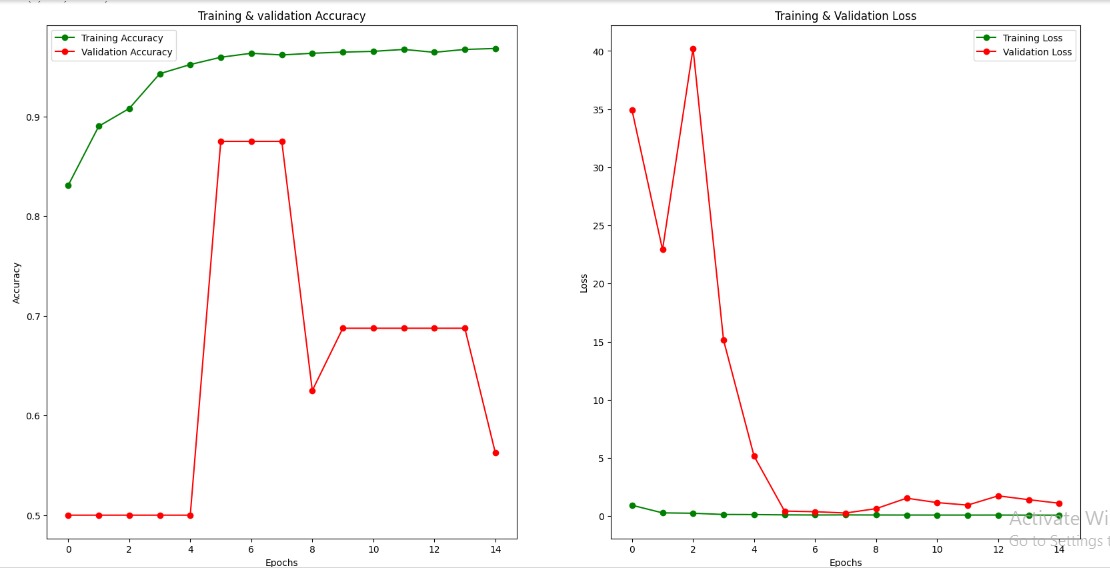
**Accuracy:** The overall percentage of correctly predicted images.

**Precision:** The ratio of true positive predictions to the total predicted positives, indicating the model's ability to avoid false positives.

**Recall:** The ratio of true positive predictions to the total actual positives, indicating the model's ability to capture true positives.

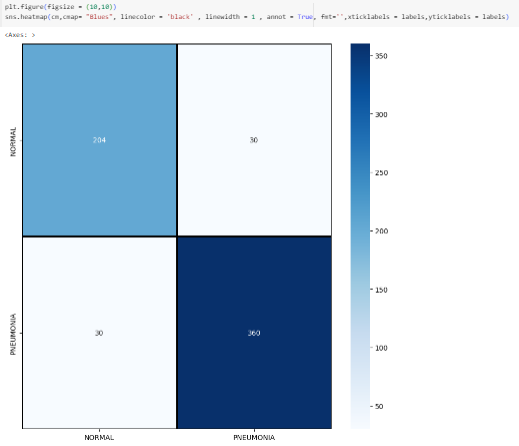
**F1-Score:** The harmonic mean of precision and recall, providing a balance between the two metrics.





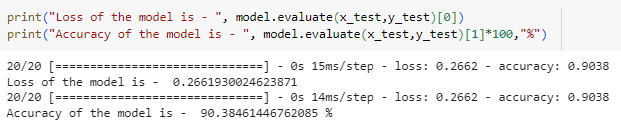
**8.2 Confusion Matrix**

A confusion matrix was generated to visualize the model's performance across different classes ("normal" and "pneumonia"). The matrix provides insights into true positives, true negatives, false positives, and false negatives, enabling a comprehensive understanding of the model's strengths and weaknesses.



**8.3 Accuracy Achieved**

The model's accuracy on the test dataset was a key performance indicator. The achieved accuracy percentage reflects the model's ability to accurately classify chest X-ray images into the appropriate categories, indicating its reliability in pneumonia detection.

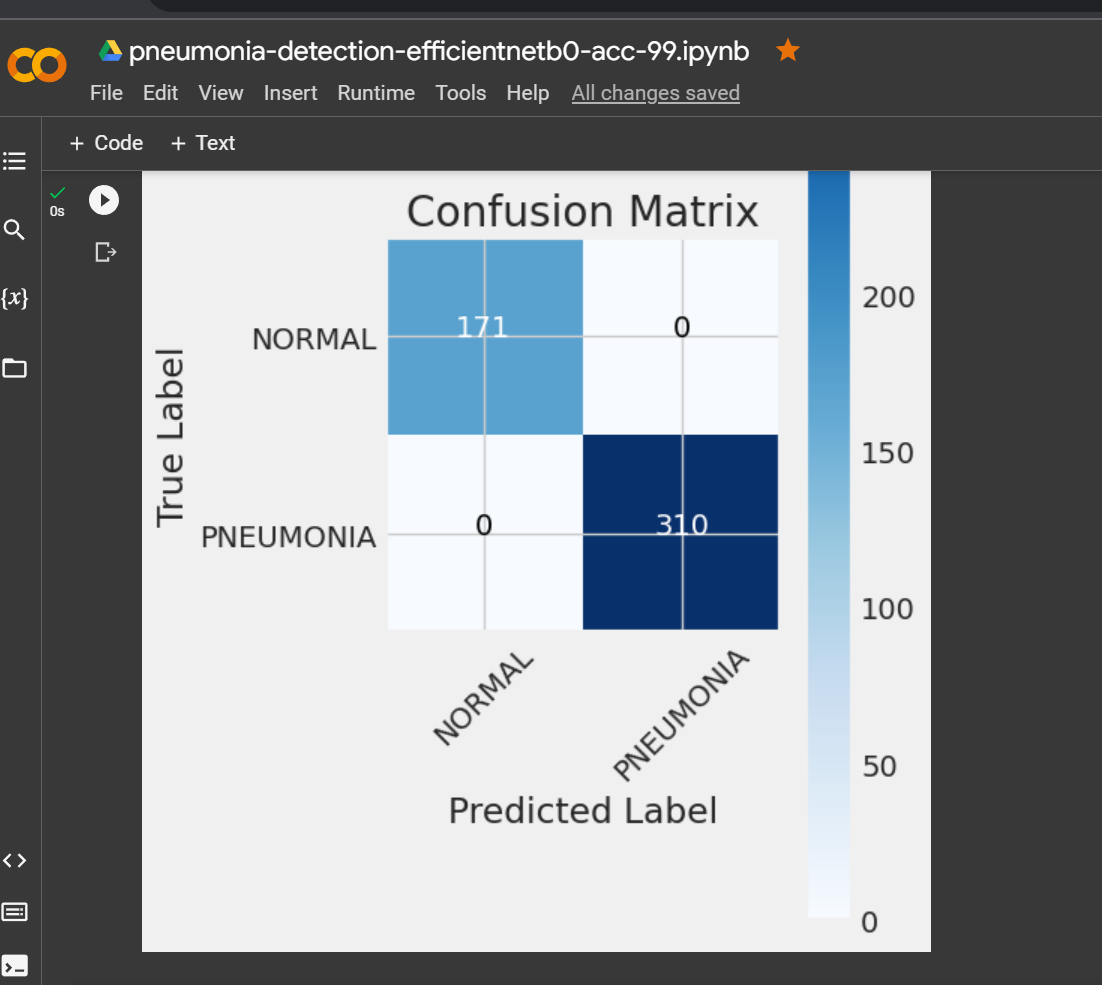


**8.4 Comparative Analysis**

We have tried some other models also for doing the prediction. First we tried the simple SVM model using sklearn library, it was giving 99% training accuracy, but was giving only 67% accuracy on the testing set, which clearly shows that it was overfitting.

Then we tried VGG16 as base model, with Sequential CNN model, but the accuracy was just around 70% which is not good for x-ray image classification.

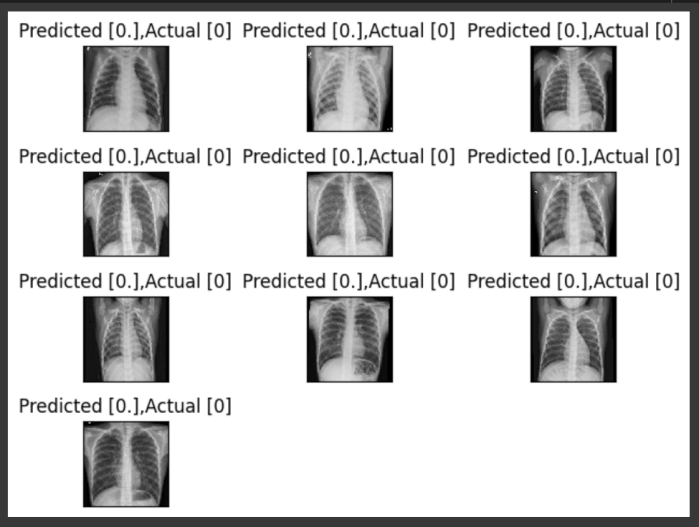
We also tried Sequential CNN with pretrained model effecientnet, tested it for 10 epochs, it was giving very good accuracy on all the sets. Below is it’s confusion matrix:

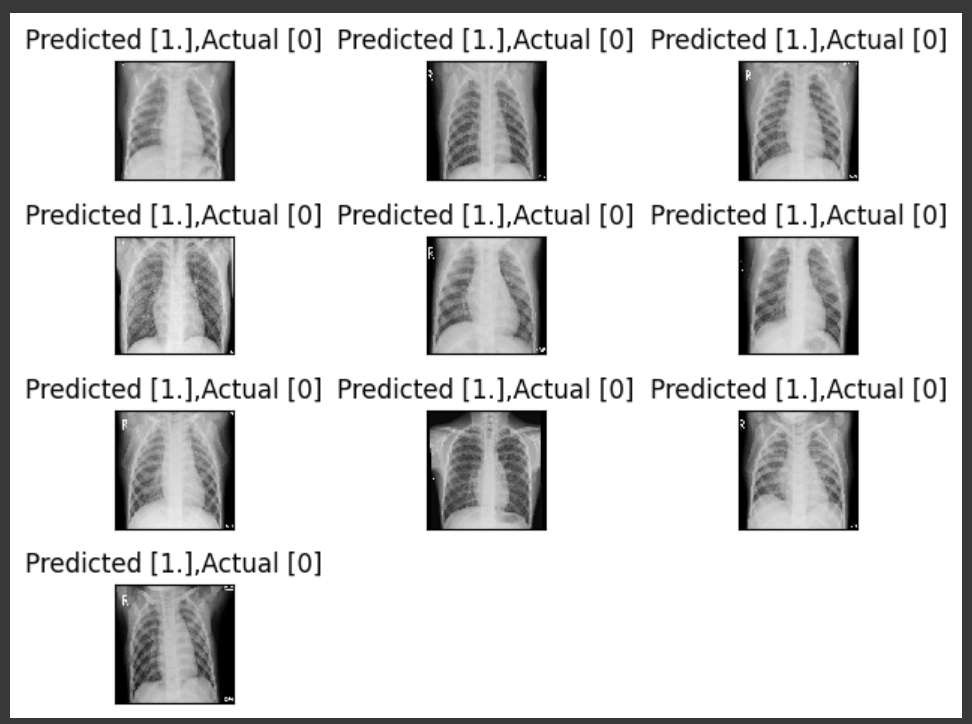


As per it’s results, it should not predict image wrongly, but unfortunately it was predicting all the “Normal” images as “Pneumonia”, for pneumonia images, it was working fine. It shows that it was biased towards Pneumonia images, so we rejected it as well.

After trying all the models, the Sequential CNN model with rmsprop optimizer and binary\_crossentropy loss function gave us the best accuracy among all, hence we saved that model for our prediction.

**8.5 Sample Predictions**





**8.6 Discussion of Results**

Overall, we have achieved an accuracy of approximately 93%, the model is able to predict most of the images correctly. We have tried and tested it with various models, hypertuned it with various parameters and added different layers to get the best accuracy among the models we tried.

**The model can be adapted for scalability, better convergence, and better accuracy.**

**8.7 Future Enhancement**

Although we have achieved 90% accuracy score on the test set, but as we are making this model for disease detection, so we have to achieve more accurate results as we can’t put anyone’s life on risk.

For this our future goals include: collecting more data, using some more pre-trained models to boost the performance, check with different parameters and get the best accuracy.

**Chapter 9**

**References**

[1] Dataset :- <https://www.kaggle.com/>

[2] CNN Model :- <https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529>

[3] Cloud Database: - <https://deta.space/>

[4] Deployment:- AWS EC2 machine

[5] Ui designing :- <https://docs.streamlit.io/>

[6] Young M and Marrie T.J 1994 Interobserver Variability in the Interpretation of Chest Roentgenograms of Patients with Possible Pneumonia Arch Intern Med pp 2729–32

[7] Albawi S, Mohammed T.A and Al-Zawi S 2017 Understanding of a convolutional neural network International Conference on Engineering and Technology (ICET) (Antalya: IEEE)

[8] Goyal M, Goyal R and Lall B 2019 Learning Activation Functions: A new paradigm for understanding Neural Networks Preprint arXiv: 1906.09529

[9] Bailer C, Habtegebrial T, Varanasi K and Stricker D 2018 Fast feature extraction with CNNs with pooling layers Preprint arXiv: 1805.03096