Automated Radiology Report Generation on Chest X-Ray

A thesis submitted by

Abdul Rafeh (GL)	**(20CS059)**
Rajnesh Charan	**(20CS053)**
Abdul Wajid	**(20CS079)**

Supervised by

Dr. Bushra Naz

Submitted in the partial fulfillment of the requirements for the degree of
Bachelor of Computer Systems Engineering

Faculty of Electrical, Electronics & Computer Engineering
MEHRAN UNIVERSITY OF ENGINEERING & TECHNOLOGY, JAMSHORO

November 2024

CERTIFICATE

This is to certify that "**Automated Radiology Report Generation on Chest X-Ray**" is submitted in partial fulfillment of the requirement for the degree of Bachelor of Computer Systems Engineering by the following students:

Abdul Rafeh (GL)	**(20CS059)**
Rajnesh Charan	**(20CS053)**
Abdul Wajid	**(20CS079)**

____

Supervisor

Dr. Bushra Naz

____

(Chairman, Department of Computer Systems Engineering)

Date: __________________

ACKNOWLEDGEMENT

Firstly, and foremost, I would thank Allah (SWT) for the opportunity to achieve this project and thesis. His grace empowered us by giving us strength, knowledge, and ability to achieve our aim. Without him, we would not have been able to achieve that.

I would like to thank my supervisor, **Dr. Bushra Naz**, with whom I conducted my Final Year Project and wrote my thesis. Her patience, motivation, and immense knowledge have been invaluable in both the development phase and while writing. I could not ask for a better supervisor at this undergraduate level.

I am thankful to Chairman **Dr. Shahnawaz Ali** Talpur from the Department of Computer Systems Engineering for believing in our idea for the project and supporting it to see the dreams transform into reality. Also, thanks for guidance and facilitation for any kind of problem my colleagues and I sought with their wise advice in various critical matters.

Last, but not least, we are grateful to our teachers, family and friends for their cooperation and help in all respects during the study period which enabled us to progress expeditiously.

ABSTRACT

The growing incidences of respiratory diseases such as pneumonia, tuberculosis, and COVID-19 have made efficient and accurate diagnostic tools important in the health sector. Current radiology practices are generally based on manual interpretation of chest X-rays (CXRs), which is labor-intensive and vulnerable to human error. This paper outlines an automated radiology report generation system based on deep learning and natural language processing to solve the problems outlined above.

The system integrates CNNs with the latest architectures like EfficientNet, ResNet, and VGG in order to classify CXRs and ascertain the presence of the three target diseases to have a robust and generalizing model, the training will be performed on a varied dataset derived from public libraries such as Kaggle. Techniques for data preprocessing-augmentation and normalization-improve the performance of the model. The best one with regard to accuracy, precision, recall, and F1-score is deployed with GPT-4 for generating reports automatically.

The application is built in Python and hosted on a user-friendly interface developed using Gradio. This enables patients to upload CXRs, view disease predictions, and download detailed radiology reports. The effectiveness of the system is validated by confusion matrices, learning curves, and qualitative analysis of the generated reports.
This solution indicates the potential of introducing AI into medical workflows by showing its scalable, accurate, and time-efficient diagnostic support capability. Future improvements might involve extending the scope of the classification of diseases, enhancement of explainability, and integration of the system with hospital information systems for enhanced clinical decision-making processes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence.
CXR	Chest X-rays.
TB	Tuberculosis.
SDGs	Sustainable Development Goals.
GPU	Graphics processing unit.
NLP	Natural Language Processing.
API	Application Programming Interface
CNN	Convolutional Neural Network.
RNN	Recurrent Neural Network.
VGG	Visual Geometry Group.
ResNet	Residual Neural Network.

CHAPTER 1 INTRODUCTION

Increased demand for healthcare through AI-based solutions has resulted in tremendous improvement in diagnostic technology. Diagnostic medicine, primarily in a critical area is radiology, which acquired much mileage due to the introduction of AI. CXR is a common diagnostic tool widely employed for diagnosing various kinds of respiratory illnesses, which include pneumonia, TB, and COVID-19 [1]. TB is still among the leading ten causes of death globally with a projected 10.6 million cases in 2021 alone [2]. Likewise, pneumonia and COVID-19 have proved to be a huge challenge to health care systems everywhere, and the need for timely and accurate diagnosis can never be overemphasized [3].

Traditional radiology reporting techniques are largely manual, demanding substantial time and expertise from the radiologists. Increasingly, radiologists find it challenging to produce reports within a timely and accurate manner, thus delaying diagnoses [4]. Moreover, due to the variability between observers in intra- and inter-observer variability, manual reporting can easily compromise the reliability of diagnosis, especially in resource-limited settings [5].

This project, "**Automated Radiology Report Generation for Chest X-Rays**," addresses these limitations as it develops an AI-driven application for automated radiology reporting. The application detects three major respiratory conditions on chest X-rays, focusing on pneumonia, tuberculosis, and COVID-19 in particular, using deep learning-trained models on annotated datasets. The automated solution reduces workloads on radiologists to increase efficiency and accuracy levels in diagnosis. The system, by using CNNs and transformer-based models, will give a detailed standard diagnostic report with recommendations for further clinical evaluation.

The project will be very important to areas with low accessibility to radiology knowledge and expertise, as this will provide rapid and low-cost diagnostic solutions. With this system, there's potential to make the workflows of radiology smoother and help patients get better outcomes through efficient care from healthcare providers.

1.1 BACKGROUND

Before delving into the proposed methodology, it is crucial to understand some foundational concepts related to chest X-rays and AI-driven diagnostic systems.

1.2 CHEST X-RAYS IN MEDICAL DIAGNOSIS

Chest X-rays are available imaging tests that create images of the internal structures of the chest, including the heart, lungs, and bones, by using low levels of ionizing radiation. They are extensively applied for diagnosing several pulmonary and cardiac conditions because they are widely available, inexpensive, and non-invasive. However, interpreting the images produced in a CXR accurately requires specialized training and experience; thus, AI solutions are increasingly invaluable for bridging the diagnostic gap in underserved areas [6].

1.2.1 PNEUMONIA, TUBERCULOSIS, AND COVID-19 DETECTION

Pneumonia is characterized by inflammation within the air sacs in the lungs, which might be caused by bacteria, viruses, or fungi. Tuberculosis is a bacterial infection caused by Mycobacterium tuberculosis; it primarily affects the lungs but spreads to other parts of the body as well. COVID-19 is a virus caused by SARS-CoV-2. Symptoms include respiratory distress, and it is much worse when there is serious damage to the lungs. Early diagnosis of these cases is very crucial in averting the high mortality rate and stopping the spread [7, 8].

1.2.2 DEEP LEARNING IN MEDICAL IMAGING

Deep learning techniques, particularly CNNs, have revolutionized medical imaging by enabling automated feature extraction and pattern recognition. Models such as ResNet and DenseNet have shown remarkable performance in disease classification tasks, while transformer-based architectures like Vision Transformers (ViTs) are increasingly adopted for improved accuracy in complex diagnostic scenarios [9].

1.2.3 AUTOMATED REPORT GENERATION

Automated reporting in radiology creates formatted diagnostic reports from the original imaging data. It makes it error-free, non-variable, and also speedier. It uses NLP with image analysis for conversion of image findings into rich textual detail that can support health providers in healthcare decision-making processes [10].

1.2.4 REAL-TIME IMPLEMENTATION

Deployment of AI models to undertake real-time analysis and report in clinical environments forms an essential part of this project. Real-time functionality permits instant feedback for diagnosis; in this way, clinicians could activate appropriate treatment plans promptly while saving more patients.

1.3 MOTIVATION

This project is based on the critical importance of accurate and timely radiology reporting to improve patient outcomes and enhance healthcare efficiency throughout the world. Diseases, such as pneumonia, tuberculosis, and COVID-19, have been a problem for public health, mainly due to high morbidity and mortality rates, even in resource-limited setups. Manual interpretation of chest X-rays is time consuming, error prone, and usually dependent on the availability of experts in radiology.

This thesis intends to use the power of artificial intelligence to address the above challenges by automating radiology report generation. By integrating deep learning models into the diagnostic process, this project aims at making radiology reports more accurate, consistent, and quicker so that healthcare professionals may focus on critical decision-making and patient care.

This is the aspiration that motivates such an initiative, to make advanced health care technologies available and scalable, especially for the benefit of underserved regions, to contribute toward the wellbeing of patients, professional growth of healthcare providers, and overall improvement in the systems of healthcare delivery in Pakistan. Beyond innovation in terms of technology, this is a commitment toward addressing the real-world health care problems and advancing the bigger goal toward equitable health care for all.

1.4 PROBLEM STATEMENT

The interpretation of chest X-rays is imperative in the diagnosis of many critical diseases, such as pneumonia, tuberculosis, and COVID-19, against which the world is under a significant threat. Globally, healthcare systems still face critical issues in being able to meet the growing demand for radiological diagnosis because of a lack of skilled radiologists, predominantly in underserved and remote regions. This results in delayed diagnosis, inconsistent reporting, and diagnostic errors. These factors affect patient outcome and burden healthcare providers.

Manual interpretation of chest X-rays is time-consuming, susceptible to human error, and difficult to scale in the face of rapidly increasing data volumes, especially during global health crises such as the COVID-19 pandemic. Such limitations underscore the pressing need for innovative and automated solutions to support radiological diagnostics.

This project thus presents an automated radiology report generation system based on deep learning, which helps in overcoming the challenges stated above on a global level.This solution targets reducing diagnostic delays, accuracy, and better decision making by radiology professionals for patient care all over the world with the right diagnostic capabilities being accurate, consistent, and scalable. In addition, it lays a roadmap to integrate AI-based methodologies in the practice of radiology for improved health care delivery efficiency and access.

1.5 AIM AND OBJECTIVES

Aim:

Development of a high-end automated system, incorporating deep learning and AI, which can produce accurate, high-resolution radiology reports for chest X-ray images, thus aiding in accurate diagnoses of pneumonia, tuberculosis, and COVID-19.

Objectives:

Dataset Collection: Collection of different images of chest X-rays representing all types of conditions, such as pneumonia, tuberculosis, COVID-19, and normal conditions, with high detail to train the model thoroughly.

Data Preprocessing: Perform required pre-processing methods such as removal of noise from X-rays, normalization, and feature enhancement to optimize the quality of the X-rays, while at the same time pulling out more relevant features.

Model Construction: Such a deep model was created and trained for reliable classification of diseases, to be highly accurate across the chosen diseases.

AI Integration: The output of the disease classification should be integrated with an OpenAI API to automatically generate professional, human-readable radiology reports.

System Validation: Test and evaluate the system's performance on unseen data, measuring accuracy, precision, recall, and overall reliability in real-world diagnostic scenarios.

Scalability and Usability: Scalable for deployment and user-friendly make the system efficient to adopt in healthcare environments globally.

1.6 SCOPE

This project will focus on specific, detailed, and professional reports through the incorporation of OpenAI APIs into deep learning models that present an application globally and are the efficient solution for the generation of automatically produced radiology reports concerning chest X-rays with special interest in diagnoses of pneumonia, tuberculosis, and COVID-19. This presents major issues in radiology as there is a significant lack of qualified radiologists in underserved areas. Other major problems are human errors, which might also cause delayed diagnosis or possible misdiagnosis. The production of reports through automation greatly helps to reduce the burden on the doctors, improve care, and hasten decisions that help diagnose and further the treatment process. Since the project is scalable and modular, it can quite easily be extended to medical imaging fields beyond chest X-rays. This has openings to offer wider applications within healthcare and has contributed more to refined diagnosis, the reduction of costs in healthcare, and greater accessibility toward quality healthcare for a world of people.

1.6.1 SDGS COVERED BY THE PROJECT

1.6.1.1 SDG 3: GOOD HEALTH AND WELL-BEING:

- This will improve health care through appropriate and timely diagnostics by improving patient outcomes and supporting care systems for health care professionals in less developed areas.

1.6.1.2 SDG 9: INDUSTRY, INNOVATION, AND INFRASTRUCTURE:

- This project, then, with deep learning and AI incorporation, shows innovations in radiology and, thus, the application of advanced technologies in the improvement of world health infrastructures.

1.6.1.3 SDG 17: PARTNERSHIP FOR THE GOALS:

- This project underlines the necessity for collaboration between healthcare facilities, radiologists, and the developers of technology for perfect use of AI-driven diagnostic tools. Skilled radiologists are required for system verification and maintaining the level of clinical standards. Clinical centers also require well-crafted, high-quality images of radiology to attain optimal performance. All this collaboration will strengthen health care services and shared knowledge leading to a more effective and reliable diagnostic solution to support patients and providers.

1.7 COMPLEX ENGINEERING PROBLEM:

Creating Scalable and Interpretable Automated Report Generation System for Clinical Deployment to be Developed from Chest X-rays

1.7.1 PROBLEM STATEMENT

Design and implement scalable, explainable, clinically deployable automated chest X-ray report generation system that captures a real-time processing mechanism to classify images into the following categories with high accuracy as well as robustness across patients' diversity: Normal, Pneumonia, Tuberculosis, and COVID-19. Produces clinical radiology reports with the help of an NLP model which follows clinical guidelines and terminologies.

The key challenges which the solution has to address include the following:

Variability in data and its quality: Different image qualities, resolutions, and noise within different datasets.

Performance and generalization: Consistently high accuracy and generalized ability to work with data that may be imbalanced and sparse like tuberculosis or COVID-19.

System integration and scalability: The module needs to integrate classification and reporting-generation into one system with the aspect of low latency, considering it is for clinical, real-time applications.

Explainability and Trust: Provide interpretable outputs, such as visualizations of the location of a disease using Grad-CAM, for building trust with radiologists and sound decision-making.

Resource Optimization: Optimize the system with a small footprint for computing and memory to be deployable on resource-constrained environments such as hospitals and clinics' edge devices.

Compliance and Validation: Compliance should include medical data privacy such as HIPAA and GDPR while strictly clinical benchmarks and peer review validations

1.7.2 ENGINEERING CHALLENGES

Model Optimization: Improve the classification accuracy of deep learning models such as VGG16, ResNet50, EfficientNetB0 while reducing their computational complexity for efficient inference on large datasets.

Quality enhancement of reports: Tweak GPT-4-based radiology report generation towards diversity of medical vocabularies as well as coherence with their meaning, medical commonsense and accuracy.

Multimodal data fusion: Take on patient history or clinical indicators within the pipeline so as to increase diagnostic correctness while further improving contextual relevance for reports being generated.

Scalability: Develop distributed architecture which should help scale to thousands of processing concurrently across several facilities dealing in chest X-ray imaging.

Explainability Tools: Develop strong visual explanation tools, such as Grad-CAM, to provide radiologists with key regions of interest in the X-ray images.

Validation Framework: Develop an exhaustive testing framework using benchmark datasets, for example, MIMIC-CXR, IU X-ray, to assess clinical accuracy, natural language metrics, and system robustness.

1.7.3 EXPECTED OUTCOMES

1. A fully functional automated system capable of real-time classifying and reporting on chest X-rays.
2. Quantifiable improvements in both clinical metrics such as accuracy, precision, and recall and textual metrics like BLEU, ROUGE, and CIDEr.
3. A deployment-ready solution with low resource requirements: good for real-world application in clinical settings.
4. Boosting trust and usability, explained AI, and conforming to medical standards.

1.8 THESIS LAYOUT

This thesis is a conceptual outline of various chapters detailing the different aspects of research and development carried out in the automated chest X-ray report generation system. These are:

- Chapter 2: Literature Review
  Reviews the current approaches for interpretation of chest X-rays in pneumonia, tuberculosis, and COVID 19. Presents the gaps so far identified by this project.
- Chapter 3: Methodology
  It explains dataset collection, preprocessing, model development, and system integration for the generation of an automated report of diseases.
- Chapter 4: Results and Discussion

Presents the system's performance metrics, detection accuracy, report quality, strengths, and limitation.

- Chapter 5: Conclusion and Future Recommendations

  Summarize the findings with their relevance to health care and recommendations for improvement in the future, such as disease coverage and accuracy.

CHAPTER 2 LITERATURE REVIEW

The last few years have witnessed huge attention on the generation of automatic radiology reports, due to their ability to facilitate the timely and accurate determination of disease from medical images by a radiologist. Deep learning models, CNNs especially, have impressed everyone by delivering remarkable results in classification tasks involving image recognition: the detection of diseases like pneumonia, TB, and COVID-19 from chest X-ray images.

2.1 Deep Learning Models for Chest X-ray Disease Classification

The CNN has emerged as one of the leading deep learning architectures in the classification of chest X-rays. A lot of studies show the effectiveness of CNN architectures in identifying and diagnosing diseases using chest X-ray images.

A study by Rajpurkar et al. (2017) applies CNN model to detect pneumonia from chest X-rays where they reported that the CNN model performed at the par of radiologists in the recognition of pneumonia cases [11]. The model is trained from a large set of chest X-rays, so it can identify high accuracy patterns that reflect pneumonia. Similarly, in a study by Liu et al. (2020), a CNN-based model is applied to detect COVID-19 and pneumonia on chest X-rays, thereby showing high classification accuracy; it indicates the possibility of using CNNs in respiratory diseases diagnosis [12]. The use of ResNet for the analysis of medical images is another remarkable approach. Indeed, due to its much deeper network architecture, it outperforms other structures in numerous tasks, even including medical image classification. A modified version of ResNet was utilized by Liu et al. (2021) for the identification of COVID-19 using chest X-rays with more than 90% accuracy; hence, diagnosis speed and reliability have significantly improved in the clinical field [13].

2.2 EFFICIENTNET AND VGG FOR DISEASE DETECTION

Other CNN architectures which have been explored for the automated classification of chest X-rays include EfficientNet and VGG. EfficientNet has achieved high accuracy with fewer parameters and, hence, is a computationally efficient choice for medical image analysis. Tuli et al. (2020) used EfficientNet to classify COVID-19 from chest X-rays and compared the performance of this model with others such as ResNet and

VGG. Their results indicated that EfficientNet surpasses both in terms of accuracy and computational efficiency [14]. Another application of the VGG model, which is quite simple and easy to implement, is in medical imaging. Vasquez et al. (2019) applied the VGG-16 architecture for classifying chest X-rays into different categories, including normal and pneumonia. This performed very well with higher costs, though, especially compared with a much more up-to-date architecture like the EfficientNet [15].

2.3 TRANSFER LEARNING FOR MEDICAL IMAGE CLASSIFICATION

Transfer learning has become widespread in the medical domain for fine-tuning a pre-trained model to a medical imaging task. This transfer learning approach is used on large datasets such as ImageNet to train models and then, applies them directly to smaller, specialized datasets in medicine.

Chouhan et al. (2020) applied transfer learning in detecting pneumonia in chest X-rays. Through fine-tuning a pre-trained VGG model, the authors were able to attain very high accuracy given the limitation of the used dataset; this demonstrates well the usefulness of transfer learning in tasks of medical images classification [16]. Especially when dealing with data scarcity for which there's not a lot of labeled images, like in medical ones, the use of pre-trained models is really practical.

2.4 GPT-3 FOR AUTOMATED REPORT GENERATION

Generating an image description report is achieved post the disease identification. With traditional means, the generating of reports was always something that required the efforts of a human radiologist with time consuming and sometimes an error in human nature but integrating GPT4 for report generation has come out very promising.

For instance, in a study published by Yang et al. (2021), the GPT-3 was leveraged for automated radiology report generation. In such a case, the model produced textual descriptions of chest X-ray images based on the detected conditions, say pneumonia, TB, or COVID-19. Here, the authors concluded that GPT-3 produced coherent and relevant reports very close to human-written descriptions, which makes it a viable tool for assisting radiologists [17].

Based on this, GPT-4 is even much better suited for the use case of automatic generation of medical reports, given improved language capabilities. GPT-4 can understand tough medical terminology and generate very accurate reports within context because its practical application in using AI in radiology enhances the speed and accuracy associated with diagnosis, especially if made in areas where resources may be lacking and entry by radiologists is confined.

2.5 LEVERAGING WARM STARTING TO IMPROVE THE CHEST X-RAY REPORT GENERATION REPORT

Warm starting has turned out to be another very promising but under-explored approach for generating reports from CXR. This approach involves initializing a model's parameters using those of another pre-trained model which have been trained on some pre-training task.As noted above, this enables the transfer of much learned knowledge about the patterns from the pretraining data that will improve performance on the target task. This technique is especially helpful when the size or quality of the target dataset is low and the task for the pretraining belongs to the same domain as well as if the size or quality of the pretraining dataset is high [20].

2.6 AUTOMATED CHEST X-RAY REPORT GENERATOR USING MULTI-MODEL DEEP LEARNING APPROACH

We proposed in this study a multi-model deep learning-based automated CXR report generation system to support the radiologists in improving diagnostic efficiency and accuracy. It was specifically designed to detect abnormalities in CXR images followed by the generation of reports accordingly, based on clinical standards. Unlike previous research activities [4]–[8], which focused merely on the detection or classification of a single abnormality in a radiological image, the proposed research expands the horizon to detect multiple abnormalities existing in one image.

A multi-model framework is being used. Each deep model is specifically designed to identify its own abnormalities, such as pneumonia, tuberculosis, and COVID-19.This methodology ensures complete analysis by providing the capability to identify the co-existence of diseases, which would otherwise be missed. This system, while improving diagnostic accuracy, allows the integration of these models into a smooth mechanism for the generation of automated reports. This solution is most helpful in resource-poor environments, where expert radiologists are not accessible and diagnostic ability and patient care are improved.[10]

2.7 DEEP-CHEST: MULTI-CLASS DEEP LEARNING MODEL FOR CHEST COVID-19, PNEUMONIA AND LUNG CANCER DIAGNOSIS

Deep Learning has been highly effective for creating models that would effectively predict and classify multiple conditions through medical images. Such studies involved the following diseases: breast cancer [6], liver disorders [7], colon carcinoma [8], brain tumor diseases [9], skin carcinomas [10], pulmonary carcinomas [11], pneumonias [12] and recently COVID-19 disease. The main strength of deep learning is that it can automatically learn features by creating abstract representations of data as the network goes deeper, unlike traditional machine learning, which requires hand-crafted feature extraction. Deep learning techniques optimize feature learning through a combinatorial arrangement of a series of nonlinear functions to maximize model accuracy without human intervention.[21]

2.8 CHANGES IN THE BRONCHIAL CUFF PRESSURE OF LEFT-SIDED DOUBLE-LUMEN ENDOTRACHEAL TUBE BY LATERAL POSITIONING: A PROSPECTIVE OBSERVATIONAL STUDY

The current research attempts to find the influence of positional changes on the pressure of bronchial cuffs in thoracic surgery patients. A majority of the thoracic surgeries require double-lumen endotracheal tubes for achieving one-lung ventilation. Suitable cuff pressure should be maintained without creating air leaks so as to ensure surgical precision along with preventing complications like trauma or mucosal edema to the airways. This research focused on the effect that occurs when a patient's position is changed from the supine to the lateral decubitus during common thoracic procedures.

The authors recruited 69 patients aged 18–70 years and measured BCP at different intervals before and after transitioning to the lateral position. This indicated that the BCP at statistically significant increased from 25.4 ± 9.0 cmH2O in the supine position to 29.1 ± 12.2 cmH2O in the lateral decubitus position with a p value of less than 0.001.There was also clear asymmetry observed in increases, more marked with LLD than with the RLD group. This observation suggests that body positioning influences cuff pressure due to tracheal morphology and the gravitational effects on the DLT.

The study highlights the importance of monitoring and adjusting BCP to avoid complications, especially since DLTs are stiffer in nature and larger in diameter than single-lumen tubes. The study also highlights the importance of fiberoptic confirmation of correct placement of the tube to avoid pressure changes due to displacement. Although the findings are specific to DLTs, they add valuable insight into the broader implications of cuff pressure dynamics in airway management.

This research study is relevant to our project as it stresses the importance of embedding domain-specific issues within the automated systems. Such clinical nuances in the interpretation of our chest X-ray report generation system would determine robust models aligned with radiologists' workflows. For instance, through embedding knowledge of anatomical variation along with procedural nuances, the system would produce more context-conscious and clinically accurate radiology reports. This will have a positive effect on trust toward physicians and increase the relevance of automated solutions in health scenarios.[22]

2.9 EFFICIENTRANSNET: AN AUTOMATED CHEST X-RAY REPORT GENERATION PARADIGM

Imaging via a chest X-ray plays an important role in the diagnosis of a wide range of diseases affecting the chest and has become indispensable in both clinical and research settings. Generation of X-ray reports might be automated to address certain limitations associated with the present manual diagnosis. It reduces reporting time, the workload for radiologists, and offers an efficient assistant that raises diagnostic accuracy. However, the problem is to capture abnormal findings with high clinical accuracy and yet produce fluent and natural language reports.

EfficienTransNet is proposed as a new approach to this challenge. It is an automatic report generation approach based on a hybrid architecture of Convolutional Neural Networks (CNNs) and Transformers. The model prioritizes the clinical accuracy of the output along with improving text generation metrics. In contrast to most previous methods, EfficienTransNet incorporates clinical history or indications directly into the generation process, which is more in keeping with the workflow of most radiologists, who utilize contextual patient information to support their interpretation of medical images.

The effectiveness of EfficienTransNet is shown on two widely used datasets: MIMIC-CXR and IU X-ray. The model demonstrates promising results in both the natural language evaluation metrics (e.g., BLEU, ROUGE, CIDEr) and the clinical accuracy benchmarks. Further qualitative evaluations, aided by Grad-CAM visualizations, highlight the specific regions of the X-ray images associated with detected abnormalities. This feature provides radiologists with critical disease location information, improving their understanding of the model's decisions.

One of the key innovations of EfficienTransNet is its focus on explainability and transparency. It emphasizes the inclusion of clinical history and visual explanations using Grad-CAM, hence fostering trust and usability in radiologists. This allows for better adoption of the system in real-world scenarios and makes sure that the generated reports are very close to the diagnostic workflow of medical professionals.

The efficiency with which TransNet bridges the gap between advanced AI technologies and practical clinical requirements sets it apart. It not only addresses the dual needs of clinical accuracy and natural language fluency but also integrates radiologists' feedback loops, making it a new benchmark for automated medical report generation. This paradigm highlights the potential of AI systems to enhance the efficiency, reliability, and scalability of radiological diagnostics in healthcare settings.[14]

2.10 CHALLENGES AND FUTURE DIRECTIONS

Despite all these achievements, several challenges still linger. The most important among them is the generalizability of models across varied healthcare settings. Models fine-tuned on specific datasets don't work well on images from different hospitals or locations. Data augmentation techniques as well as domain adaptation techniques are being explored to face this problem [18].

Furthermore, though the deep learning models have proved to be very accurate in disease detection, there remains the challenge of interpretable AI. Radiologists should know how the models reached their conclusions, which is very critical for trust and adoption in clinical practice. Explainable AI (XAI) methods are under active research to provide transparent reasoning behind AI-generated results [19].

2.11 CONCLUSION

The CNN, EfficientNet, and ResNet models showed great capabilities in the diagnosis of diseases such as pneumonia, tuberculosis, and COVID-19 from chest X-ray images. Transfer learning and GPT-4 might represent promising solutions to completely automate the detection and reporting part of radiology that would make it easier for radiologists to deliver quicker and more accurate diagnoses. However, there are model generalization and interpretability concerns that have to be addressed before being fully incorporated into clinical practices.

CHAPTER 3 DESIGN AND METHODOLOGY

3.1 OVERVIEW

This chapter covers methodology, and the design framework adopted for building an automated system to provide radiology reporting on chest X-rays, thereby employing an automated report generation mechanism. The approach takes its start from creating an elaborate dataset; following an extensive number of public datasets made available by Kaggle. Techniques of data preprocessing and data augmentation are applied next toward preparing the dataset to go into deep learning model training. This focus is on disease detection of pneumonia, tuberculosis, and COVID-19 by the application of CNN architectures like EfficientNet, ResNet, and VGG. Once the models are trained and validated, they will be incorporated into an API that will automatically generate diagnostic reports from X-ray images. The general workflow will be enabled to let an overall easy-to-use interface be available to healthcare providers, radiologists, and patients to upload chest X-ray images which will instantly prompt the report generation. It includes all methodology phases such as data collection, preprocessing, model selection, integration, testing, and finally application deployment. The overall design is to have a stable and efficient system that could help medical professionals by producing the radiology report automatically.

3.2 DESIGN AND IMPLEMENTATION

Design and Implementation of this project phase will design an automated generation system of radiology report for diseases such as pneumonia, tuberculosis, and COVID-19 from chest X-ray images. Key Components include data collection and preprocessing, model training, interaction with GPT-4 for the generation of the report, and application development of the interface easy for use.

Data gathering began by downloading a huge dataset containing chest X-ray images of pneumonia, tuberculosis, COVID-19, and normal cases from Kaggle. The dataset was divided into mainly two directories, train and test directories that contain all subdirectories of respective diseases. In total, 10,936 images were found in the training set and 2,736 images were available in the test set. To ensure that the learning and evaluation are balanced, the dataset was divided to 80% training and 20% testing. Data preprocessing was done on all images, resizing all images to a standard resolution of 196x196 pixels for uniformity.

Data augmentation techniques such as random rotations, horizontal flips, and zooming were further applied to the training set in order to increase the variability of the dataset and enhance the generalization capability of the model. These enhanced images helped the model process unseen data during training. On the test set, the rescaling was performed, and pixel values were normalized just before evaluation. In training a model, four deep learning architectures were used for the purpose of disease classification: Traditional CNN, EfficientNet, ResNet, and VGG. These models were adopted because they have a superior performance in image classification applications. The models were trained on the preprocessed training dataset, and evaluation was carried out using classification metrics such as accuracy, precision, recall, and F1-score. Upon proper evaluation, the VGG model was finalized since it had outstanding performance, with 93% accuracy on training and 92% accuracy on validation. Finally, the GPT-4 model was added to the VGG model to generate detailed radiology reports. The result from the VGG model classification of the chest X-ray image, with its predicted disease and its associated confidence score, is used and passed to GPT-4 through OpenAPI. GPT-4 now generates a radiology report that is very descriptive. It includes a diagnosis with confidence level, plus any other relevant clinical information. The reports generated will be accurate and easy to understand because GPT-4 is integrated. A web application was developed using Gradio to make the system accessible and user-friendly. The application allows users to upload chest X-ray images, which are processed by the trained VGG model for disease classification. After processing, the system shows the predicted disease and confidence score along with a downloadable PDF report containing the radiology findings.This makes the web application very simple, offering a friendly and non-techno-gloomy way to guide people who may not even know how to upload their images, letting them come through to obtain output with efficiency.

This will integrate the entire system which includes designing, that involves image preprocessing, selection of a model, report generation and much more such that user experience stays as natural and user-friendly. Through the amalgamation of deep learning with GPT-4 for generating reports, it aims at providing a totally integrated solution for an automatic disease-detection report to enhance speed and precision for chest X-ray diagnosis.

3.3 METHODOLOGY

The present paper elaborates the methodology involved in developing an automated radiology report generation system of chest X-ray images concerning pneumonia, tuberculosis, and COVID-19. It is evident that the adopted methodology is systematic and has advanced step-by-step through the following phases, such as data collection, preprocessing, training of models, disease detection, integration to generate the report, and development of an application of user friendliness. To add strength to the system to become robust, accurate, and directly applicable in health scenarios, each phase is built atop the previous one. It plans to develop a scalable, reliable system that will accurately diagnose certain diseases by interpreting chest X-rays and can autonomously produce complete radiology reports.

This is done using the state-of-the-art deep learning models, while keeping in mind that the system is accurate as well as interpretable. The methodology adopted is balanced, considering technological innovation with practical application, such that the system developed meets both the requirements of the academic and real world.

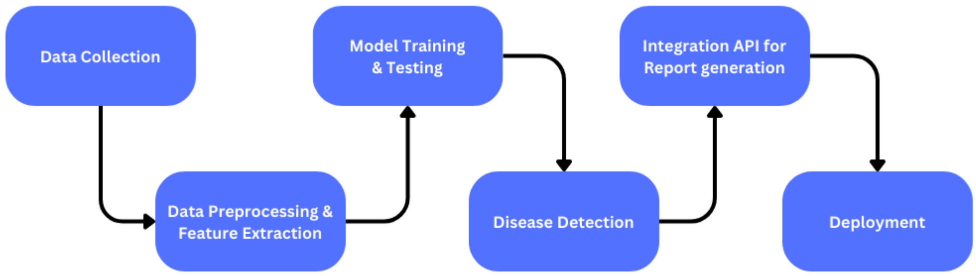

Figure 3.1 Steps of Methodology

3.3.1 DATA COLLECTION

Chest X-ray images were downloaded for this project from Kaggle, which is a prime source of medical image datasets. The dataset comprises four categories, including **covid, normal, pneumonia,** and **tuberculosis** images. These images were fetched from publicly available medical image repositories, thereby making sure the dataset was large enough to include all kinds of variations of the diseases. The dataset was split into a training set with 10,936 images and a test set containing 2,736 images. Data was arranged in respective directories for each class and further preprocessed for training.

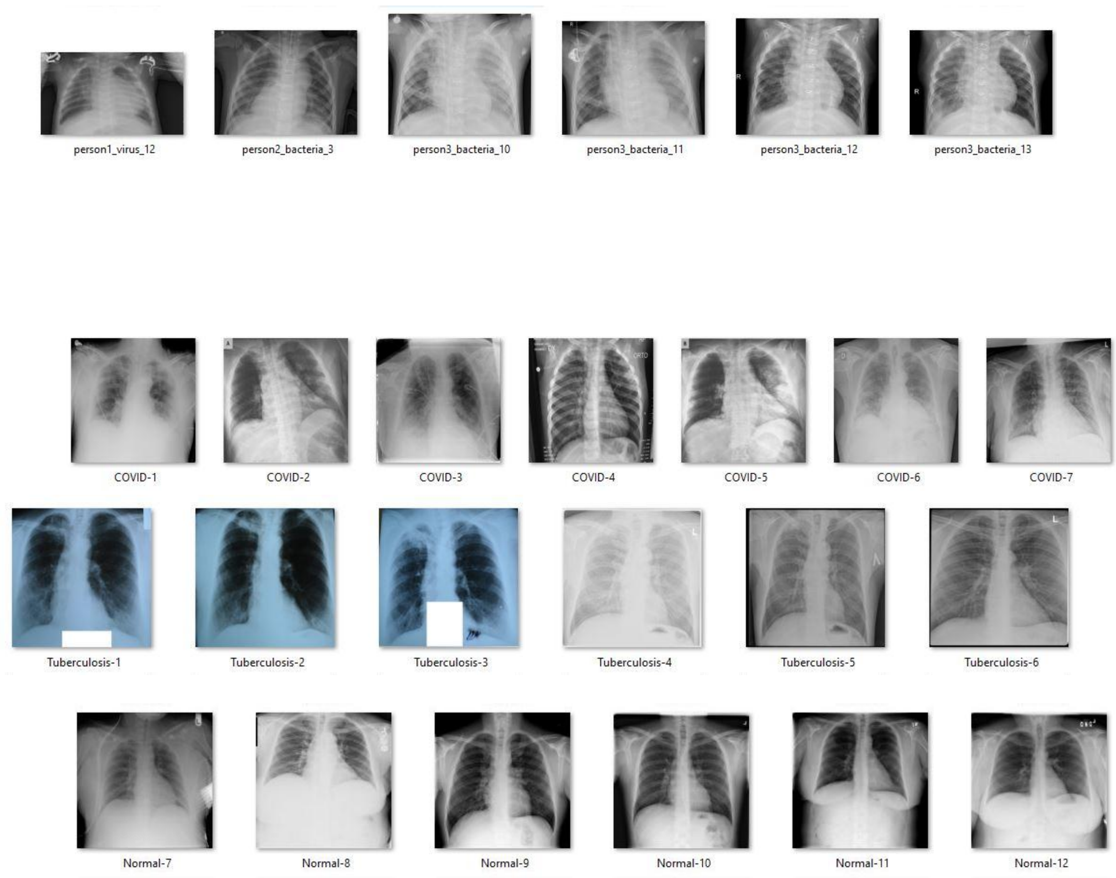

Figure 3.1 Dataset

3.3.2 DATA PREPROCESSING

Data preprocessing forms one of the critical stages necessary to prepare the data before training the model. During this stage, all the images will be presented consistently in terms of their quality, appropriately scaled and thus ready for input to deep learning models. The main tasks of preprocessing were as follows: resize all images into consistent dimensions of 196 pixels by 196 pixels and normalize pixel values between the range [0,1] and augmentation.

3.3.3 DATA AUGMENTATION

Because some of the classes were relatively small in size and because the model needed to be more robust, data augmentation was applied on the training set. These included random rotations, flips, height and width shifts, shear, and zoom, among others. These augmentations artificially enlarged the dataset, letting the model learn from a wider variety of image variations and thereby improving its ability to generalize. Data augmentation ensures that the model is not prone to overfitting and can predict well on test data.

3.3.4 DATA NORMALIZATION

Image normalization was applied to standardize the input data and make sure the models receive appropriately scaled data. All the pixel values were rescaled to a range between 0 and 1 by dividing by 255, ensuring the images are all on a consistent scale so that the neural networks may converge faster during training, improving the overall performance.

3.3.5 HANDLING DATA IMBALANCE

Given the unbalanced nature of medical images, where some conditions are underrepresented, such as tuberculosis, a weighted loss function was used to take care of this problem. Class weights were calculated considering the frequency of each class in the dataset, which were then used in training to prevent the model from being biased toward the class that occurred more frequently in the dataset. This technique allowed the model to focus on the identification of less common diseases, like tuberculosis, without compromising performance with more common conditions, such as pneumonia.

3.3.6 DATA SPLITTING

The dataset was divided into training and testing sets with an 80:20 ratio, meaning 80% of the images were used for model training and 20% for testing. This split is critical for evaluating the model's generalization ability and ensures that the model was trained on one set of images and evaluated on another unseen set to check for overfitting.

3.3.7 ALGORITHM PERFORMING

This work depends very much on deep learning classification models classifying chest X-ray images, and indeed, it used four recent advanced architectures-CNN for Convolutional Neural Networks, EfficientNet, ResNet, and VGG that were trained on the classification of various diseases such as pneumonia, tuberculosis, and COVID-19. Each model offers different strengths regarding speed, accuracy, and applicability. Listed below are detailed descriptions of each model used in this project.

3.3.8 CNN (CONVOLUTIONAL NEURAL NETWORK)

CNN is one of the most used deep learning models for image classification because it can automatically detect the spatial hierarchies within images. In this project, a traditional CNN architecture was implemented as a baseline model for disease detection in chest X-rays.

CNN models include a number of convolutional layers followed by pooling and fully connected layers. The role of convolutional layers is feature extraction, wherein edges, textures, and shapes in the X-ray images are identified. Pooling layers reduce the spatial dimensions of the data; this makes the model computationally efficient. Finally, fully connected layers classify features into the respective disease categories (pneumonia, tuberculosis, COVID-19, or normal).

This is a good starting point for the classification of diseases. The model can learn features relevant to images, but certainly not as complex or having fewer parameters than those found in more complex architectures such as EfficientNet or ResNet. The CNN model is efficient and effective for smaller datasets but lacks the depth and robustness seen in more sophisticated models.

3.3.9 EFFICIENTNET

EfficientNet is a family of models that are scaling efficiently with fewer parameters and can achieve better accuracy. Achieving better performance compared to other models in different tasks, compound scaling utilizes the uniform increase in depth, width, and resolution of a network so that better performance can be obtained using fewer computational resources.

In this project, EfficientNet was trained to detect diseases in chest X-rays. The model was first pre-trained on the ImageNet dataset and then fine-tuned on our chest X-ray data to capitalize on transfer learning. This architecture of EfficientNet consumes less resources while achieving both real-time applications and top-notch performance in terms of accuracy and speed. This makes EfficientNet a good candidate for deploying the disease detection model in resource-constrained healthcare environments, such as low-budget hospitals or clinics with limited computational resources. EfficientNet performed very well in distinguishing between various diseases, showing a high degree of accuracy in classifying chest X-rays of patients with pneumonia, tuberculosis, and COVID-19.

3.3.10 RESNET (RESIDUAL NETWORKS)

ResNet is a state-of-the-art deep learning architecture that uses skip connections, or residual connections, to enable the model to bypass the vanishing gradient problem and improve its ability to learn very deep networks. That is, ResNet lets information bypass some layers during training without losing performance; thus it is more suitable for tasks like image classification.

ResNet50, a variant of the ResNet with 50 layers, was used to accomplish the task of detection of diseases in chest X-rays. The performance by ResNet50 was tremendous because of its depth, which helped it learn complicated representations of the input images. The model learned effectively to distinguish between the diseases existing in chest X-rays with fine features of the images even when there are overlapping or subtle patterns that are usual in medical imaging. Generalizing well on unseen data is the strength of ResNet, which is quite crucial for ensuring that the model would be able to classify new chest X-ray images during deployment.

3.3.11 VGG (VISUAL GEOMETRY GROUP)

The architecture known as VGG, although very deep and has numerous layers, is recognized due to its simplicity. They make use of a pretty straightforward design that features 3x3 filters; this is how a network of such great depth will be built by just placing one convolutional filter above another. It had to be used extensively due to the high accuracy even for the relatively large number of parameters.

In this project, **VGG16** was one of the primary models for the detection of diseases in chest X-ray images. It had 16 layers: 13 convolutional layers and 3 fully connected layers. VGG16 was fine-tuned on the chest X-ray dataset with pre-trained weights from ImageNet to speed up the training process.

3.3.12 MODEL TRAINING

This stage of training is critical in the development of an accurate and reliable automated system for the detection of diseases in chest X-ray images. Four different deep learning models, such as **CNN, EfficientNet, ResNet, and VGG**, were trained on the prepared dataset for classification into categories such as pneumonia, tuberculosis, and COVID-19.

Initially, the dataset was preprocessed with normalization and augmentation to better improve model performance and generalize. Training was initiated through transfer learning for the case of the **EfficientNet, ResNet, and VGG** models by using ImageNet's pre-trained weights to expedite convergence while improving the accuracy of models. This allows the models to take advantage of features learned in large datasets that are later fine-tuned for this specific disease detection task using chest X-rays.

This has been trained using a batch size of 32 for 25 epochs in total to maintain a balance between computation efficiency and model performance. All the models applied the Adam optimizer, whose learning rate adjustment capabilities have the possibility to converge rapidly. Therefore, it's set up as an initial learning rate at 0.001, so all model training could track for the losses and the accuracies and tune for those parameters if applicable. Among these parts is preventing the most critical, overfitting and underfitting during the actual training procedure. Early stopping employed this strategy by monitoring the validation loss so that the training is stopped when the model has stopped improving its performance on the validation set. It thus preserved the generalization ability of the models without overfitting to the training data.

This task also adopted data augmentation approaches of random rotation, zooming, flipping, and shearing to artificially increase the training set size and expose the models to a wide variation of variations, hence significantly improving real-world applications of the models. Such approaches of augmentations enabled even better generalization of these models in unseen chest X-ray images under different sceneries that might conceivably occur during practical deployments.

In training, the process of the models' loss and accuracy curves over epochs helped monitor how well the learning was being done so as to prevent overfitting. Validations were executed at intervals to check models' performances on unseen data to prevent memorization over the training set but meanful patterns that could have a good generalization onto new data.

After training, the two models were tested using diverse performance metrics that compared whether they could correctly classify their chest X-rays. These metrics included accuracy, precision, recall, F1-score, and confusion matrices, which are crucial while comparing the efficacy of each model and pointing out flaws in specific models.

In summary, the training phase of a model was designed to adapt the learning of each to ensure accurate classification of the chest X-ray images while it should detect diseases with maximum accuracy, such as pneumonia, tuberculosis, or COVID-19. As the model was trained and improved in the training environment with data augmentation and monitor continuous performance, this would thus prepare for the subsequent validation and implementation of the developed models.

3.3.13 MODEL TESTING

Model testing is based on the critical evaluation of performance and reliability of models trained. In this research project, models CNN, EfficientNet, ResNet, and VGG have been tested on a set of test data that hasn't been used during training. This test dataset used includes chest X-ray images of patients diagnosed with pneumonia, tuberculosis, and COVID-19, as well as normal images. The testing set was then utilized to evaluate how well each model could classify the disease and distinguish the categories. Each trained model made predictions, which included the class label for the detected disease (pneumonia, tuberculosis, COVID-19, or normal). Some performance metrics like accuracy, precision, recall, and F1-score were used to measure the effectiveness of the models in correctly classifying the test images. Moreover, confusion matrices were generated to get a more vivid view of how well the models performed across the different classes and to highlight areas where the models probably misclassified images.

Qualitative analyses further enhanced the performance evaluation by checking predictions on test images through direct visualization. The class labels provided by the model were checked against the actual diagnoses made for the real-world situation, and the models are further compared to see if they have performed well under actual conditions. Overlap of the predicted results over the original chest X-ray images helped in observing how and where the models get their predictions correct or wrong, especially when the disease symptoms are subtle or even ambiguous.

Model testing process Cross-comparison was made about the performance of four different models to establish which one performed better in accuracy and reliability. Such cross-comparison helped bring out their strengths and weaknesses of those models in distinguishing various kinds of diseases through chest X-rays, for instance pneumonia and tuberculosis or the presence of early symptoms of COVID-19.

Finally, the testing phase of all models showed excellent performance with some variations in accuracy, precision, and recall. This ensures that the final model selected for deployment in a clinical setting will be robust and reliable for delivering accurate predictions based on chest X-ray images.

3.3.14 DESIGN AND DEVELOP APP FUNCTIONALITIES

After training and testing the model, developing the application that would let users upload chest X-ray images, receive disease predictions, and download automated radiology reports was the next step.

DEPLOYMENT OF MODEL IN OUR APPLICATION

The best selected model, integrated with GPT-4 for report generation, was deployed in a user-friendly application built using Gradio. This application allows users to easily upload chest X-ray images and receive predictions, along with detailed reports that include the predicted disease, confidence score, and additional clinical information.

APPLICATION WORKFLOW

The application workflow involves several steps to ensure smooth interaction:

3.3.14.1.1 User Interface

This is an application developed such that it caters to all levels of technical ability from the users. It has an intuitive UI design. Users can upload chest X-ray images directly either by dragging and dropping files into the system or selecting them. The system will process the image and output the diagnosis, showing the condition that the system predicted with the level of confidence attained and a detailed medical report. Moreover, the user interface provides the facility of exporting the PDF report that is generated, for easy documentation and sharing. The whole process is so smooth that it guarantees effectiveness and access by all users.

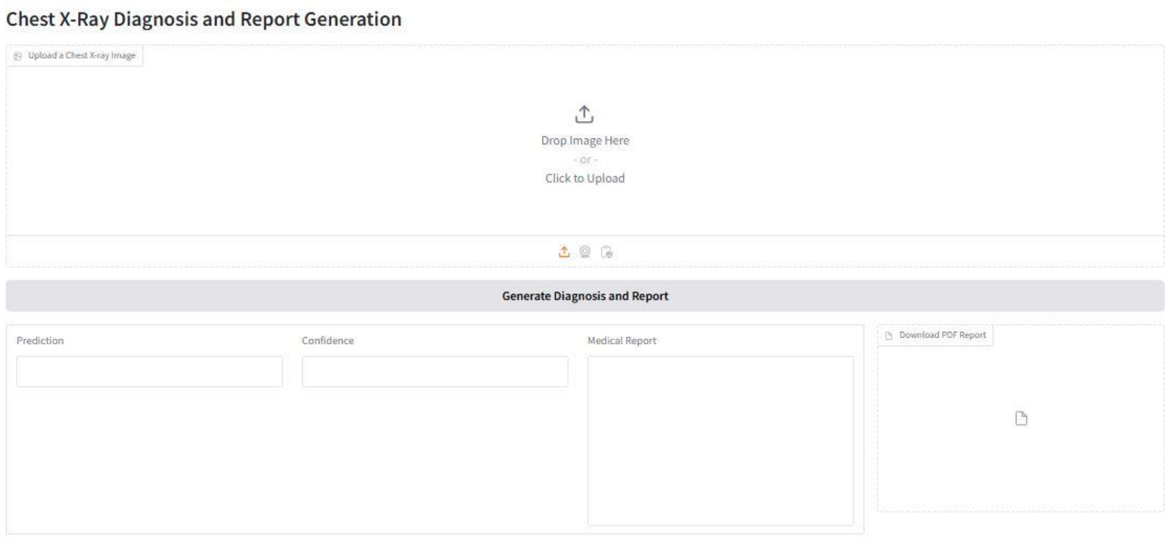

Figure 3.2: User Interface of Web application

3.3.14.1.2 Model Response

It captures the chest X-ray uploaded by the user and presents it to the chosen model, which analyzes the input further to identify and classify different cases of diseases like Pneumonia, Tuberculosis, and COVID-19. Thereafter, the application recovers the predictions from models and displays the detected conditions along with the confidence percentage and generates a detailed report for the user.

3.3.14.1.3 Display of Results

The results are easily displayed on the user interface so that easier interpretation is ensured. The chest X-ray image analysis is presented along with the prediction for the condition, which includes Pneumonia, Tuberculosis, COVID-19, or Normal. It also displays the interface along with the prediction, which describes the detected condition and its confidence level in a comprehensive medical report. This clarity in results allows easy understanding and application to decisions by users like doctors, who will respond in due time to treatment and diagnosis.

3.4 TOOLS AND TECHNOLOGIES USED FOR PROJECT DEVELOPMENT

3.4.1 PYTHON

Python has been chosen to be the main language in this project because of the flexibility and usability, alongside an impressive library ecosystem. Its syntax is simple in terms of code, so it can be used quite easily in both backend processes and model implementation in the world of machine learning. Pretty comprehensive libraries such as TensorFlow, OpenAI, and Gradio provide very good support for deep learning. A tool such as Python would be pretty inevitable for testing, training, and model deployment. Python is easy to debug and can be very effectively used for the development of applications involving machine learning

3.4.2 DEEP LEARNING

Deep learning is basically the foundation for the project. Complex neural networks make deep learning provide an opportunity for the model to learn very detailed patterns in images of chest X-rays, enabling it to make very accurate differences between diseases like pneumonia, tuberculosis, and COVID-19. Hence, it offers the ability to process enormous collections of data while generalizing the outcomes between different classes of diseases makes deep learning a good asset in developing automated radiology reports.

3.4.3 TENSORFLOW

This project employed the TensorFlow deep learning framework. The said framework is powerful and capable of creating and training models. This provides several tools and libraries for the building, optimization, and deploying of neural networks. With compatibility in Keras, it makes easy for one to create deep learning models. Deep learning requirements in architectures include CNN, EfficientNet, ResNet, and VGG;

TensorFlow supports large data processes with ease by allowing scale up, scaling down with support of acceleration in a GPU.

3.4.4 JUPYTER NOTEBOOK

Jupyter Notebook was used to interactively develop and test machine learning models. This tool offers an interactive way of writing and running Python code in cells; it is very useful in iterative testing and model improvement. In Jupyter Notebook, one can easily see the progress of training and loss functions and other metrics, hence an interactive process that makes the model better and troubleshoots the same..

3.4.5 GOOGLE COLAB

Google Colab offered a cloud-based deployment of deep learning models accessible to powerful GPU resources that significantly reduced training time while allowing for fast iterations towards model development. Google Colab supports easy collaboration for project team members by means of real-time sharing and editing of the notebooks making it a precious tool when it comes to the developing and testing stages of a project.

3.4.6 OPENAI

OpenAI's GPT-4 API was used to automatically generate radiology reports by classifying the disease using the results obtained. GPT-4 has the natural language processing ability to generate coherent, informative, and contextually accurate medical reports. Using deep learning models combined with GPT-4 helps bridge the gap between the detection of disease and clinical reporting, which allows for a smooth workflow in diagnostics.

3.4.7 GRADIO

The tool Gradio was used to create an interface that can be easily accessed for interaction with the models. It is possible to upload chest X-ray images into the system, choose the corresponding model, and receive a disease diagnosis together with a downloadable report. Gradio makes it easier to deploy the machine learning models in practical applications, allowing healthcare providers to access the system without great technical knowledge.

CHAPTER 4 RESULTS AND DISCUSSION

4.1 OVERVIEW

We present here the results of our study with an emphasis on deep learning-based approach to the detection and classification of chest diseases using images of the chest X-ray. The result of every experimental procedure undertaken here has been explained with interpretations and analysis of the outcome in order to address the major research questions of this study.

The performance of four different models such as Custom CNN, ResNet50, EfficientNetB0, and VGG16, which have been trained to predict Pneumonia, Tuberculosis, COVID-19, and Normal conditions, was surveyed in the chapter, and as demonstrated, it proved that VGG16 is the most accurate and reliable model for the task amongst the models evaluated. All the models' performance metrics will be compared against each other and presented along with their strengths and weaknesses in terms of accuracy, precision, recall, and F1 score. For further insight into the actual classification of every model, confusion matrices will be used. Discussion on how preprocessing techniques are applied on the chest X-ray images to prepare them for training and testing and in ensuring data is adequately prepared before the real training and testing. Performance of each of the reviewed models both statistically and in visual assessment of metrics showing the success of the model in its practical application. This comparative analysis underlines the better performance of the VGG16 model, and other models are discussed with their advantages in terms of higher accuracy and robustness in detecting chest diseases.

In terms of implications of these findings, the last section deals with some possible consequences in the health care sector to improve diagnostic performance to increase precision and productivity. Areas of further work: Extensions in the number of diseases covered, increase diversity of datasets, and optimization of the models.

4.2 MODEL EVALUATION

Model evaluation was carried out using several performance metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of the model’s effectiveness in detecting the various diseases. The following subsections describe how each metric was calculated and visualized.

4.2.1 PERFORMANCE METRICS

This decides the reliability and efficiency with which the machine learning models may work. In this project, several metrics have been used to give a good appraisal of the

4.2.2 ACCURACY:

Overall the model's accuracy is correct since it classified the chest X ray images correctly. This value can be calculated from a formula.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

where:

- **TP (True Positives)**:

  It denoted situations where the model **detected a diseased** X-ray to be diseased.

  Example: The model predicts **Pneumonia**, and the ground **truth** is also Pneumonia.

- **TN (True Negatives)**:

  Instances in which the model correctly classified an anomalous X-ray as **anomalous**.

  Example: Model predicts **Normal**, and also ground **truth** is Normal.

- **False Positive (FP):**

  Those are diseased X-rays that are **wrongly** identified as healthy.

  Example: As anticipated by the model- Tuberculosis. Ground Truth- Normal.

- **False Negative (FN):**

the cases on the model **misclassified** a diseased X-ray.

Example: Model is predicting Normal but actually covid-19.

4.2.3 PRECISION:

The true positive rate for each disease, relative to all of the positive predictions made by the model. It is calculated using the formula:

$$\text{Precision} = \frac{TP}{TP+FP}$$

4.2.4 RECALL:

The ability of the model to pick instances of each disease from the test set with a high degree of correctness. This can be calculated with the formula below:

$$\text{Recall} = \frac{TP}{TP+FN}$$

4.2.5 F1-SCORE:

The harmonic mean of precision and recall, providing a single metric that balances both precision and recall. It can be computed using the formula:

$$\text{F1-Score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

4.2.6 PERFORMANCE VISUALIZATIONS

The visualizations of the performance are critical in knowing the results of the machine learning model correctly. For this project, we utilized the following tools: Confusion Matrix and Precision-Recall Curves for evaluating and judging the performance of the model.

4.2.7 CONFUSION MATRIX

This generated a confusion matrix which showed the detection accuracy that each of the diseases could obtain from this model. The true positives, false positives, true negatives, and false negatives portrayed for each class would give insightful information about how strong the model was on its particular weaknesses.

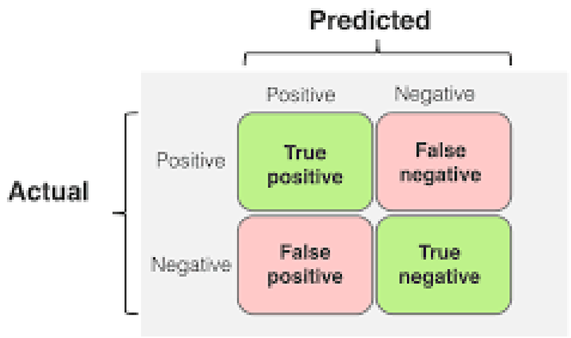

Figure 4.1 Confusion Matrix

4.2.8 LEARNING CURVES

The third part was adding plots of training epochs for loss and accuracy to monitor improvement in model performance. These curves helped to determine the overfitting and underfitting issues and guided fine-tuning to adjusted hyperparameters.

.

4.3 RESULTS OF PERFORMING ALGORITHMS

We conducted four separate experiments: Custom CNN, ResNet50, EfficientNetB0, and VGG16 for the detection of diseases from chest X-rays. Having in mind that dataset labeled for the training of chest X-ray models as "Pneumonia," "Tuberculosis," "COVID-19," and "Normal," this paper unveils the core focus of estimating and comparing those models upon accuracy, precision, and recall, and confusion matrix analysis to identify the finest solution possible for automating the detection of chest disease.

Between the models, the best was VGG16 at 92.3%. This model emerged as the most reliable to execute the given task. Then came the second best being Custom CNN at 60.0%, followed by ResNet50 at 65.4%, and then EfficientNetB0 was the poorest with 31.2% accuracy. In this case, VGG16 performed exceptionally well, as it can extract the features with complex high levels, which eventually made its performance good enough in the medical imaging cases. The lowest accuracy was by EfficientNetB0. It can also be noted that comparison of the confusion matrix depicted that classification capability for four categories of diseases was improved against other models in VGG16. Average accuracy in custom CNN and ResNet50 were very minimal, and the models were even unable to identify slight patterns of X-rays variations. This comparison clearly sets VGG16 as the best model for this application hence well-suited for its deployment in real-world automated chest X-ray diagnosis and report generation systems. Future work may surround improving other models by applying ensemble techniques so as to increase overall accuracy and robustness of the system.

4.3.1 PERFORMANCE ANALYSIS WITH CONFUSION MATRIX

The confusion matrices provide the performance of each model in classifying the four categories: "Pneumonia," "Tuberculosis," "COVID-19," and "Normal." The matrices enable us to look at true positives, false positives, true negatives, and false negatives for each class, giving insight into the accuracy and reliability of the models. In this project, confusion matrices were used to compare the ability of each model to correctly differentiate between the classes of disease. An ideal model would be able to have high true positive rates for all classes with very minimal misclassifications reflected in the off-diagonal cells of the matrix. From the matrices above, we extract specific strengths of each model and points of misclassification. Such knowledge becomes essential for improving the models, hence evaluating their readiness for deployment in real healthcare conditions. Results from this analysis included that VGG16 proved to be quite robust; it was barely misclassifying anything, thus proving highly reliable for use in automated diagnosis chest X-ray.

4.3.2 VGG16

From the confusion matrix of VGG16, it has shown very good performance in classifying the chest X-ray images into the four classes: "COVID," "Normal," "Pneumonia," and "Tuberculosis." Key points are as follows:

- COVID Class: For 640 samples, it correctly classify "COVID" and is more confused and misclassifies 74 samples with "Normal", less confounded with the other class. This, therefore demonstrates that VGG16 would very effectively classify cases of COVID-19.

- Normal Class: The model demonstrated outstanding performance for the "Normal" class, where 980 were correctly classified, and only 21 cases were incorrectly classified as "COVID" or another disease. This is an indicator of very high reliability for the detection of normal cases.
- Pneumonia Class: For the pneumonia class, 800 samples were correctly classified; however, there were 49 instances of misclassification with "Normal," and a few with "COVID." This indicates minor overlap of features between pneumonia and normal classes.
- Tuberculosis Class: The model got 110 tuberculosis cases right but was confused and misclassified 24 of them as "COVID." Although its performance on this class is good, it still needs to be differentiated from other classes.

VGG16 was at 92%, which made it the best model for this classification task. This shows the confusion matrix about how good it is in handling chest X-ray images complexity because most of the predictions tend to concentrate along the diagonal of this matrix that will give evidence about the proper classification. Therefore, it has minimal misclassifications that make VGG16 a highly reliable tool which can be deployed into the automation of chest X-ray diagnosis systems, particularly on cases of COVID-19 diseases and normal cases. There is, however, a scope for further optimization in categories for tuberculosis and pneumonia towards better accuracy in selected edge cases.

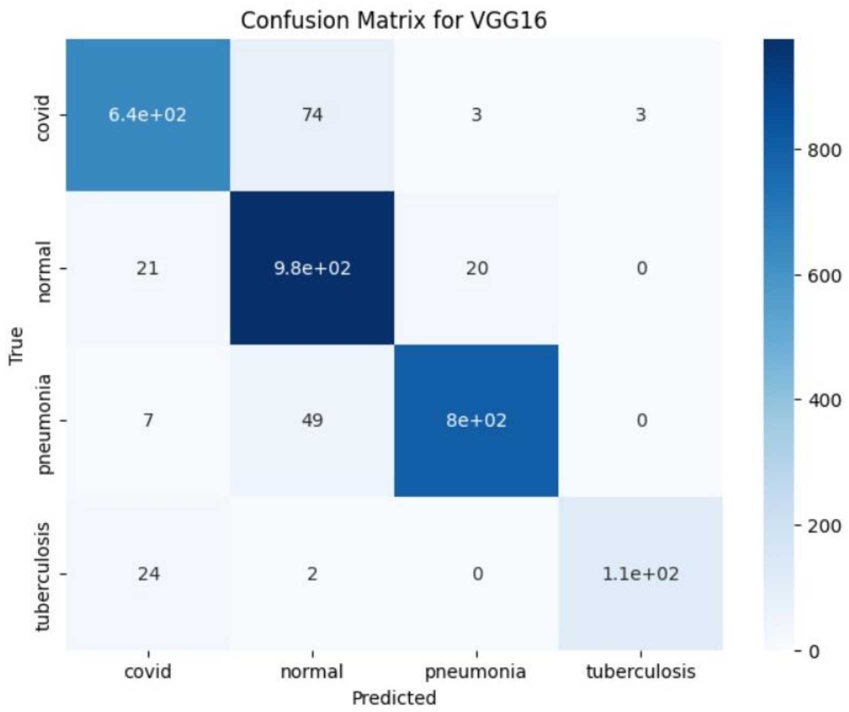

Figure 4.2 Confusion Matrix of VGG16

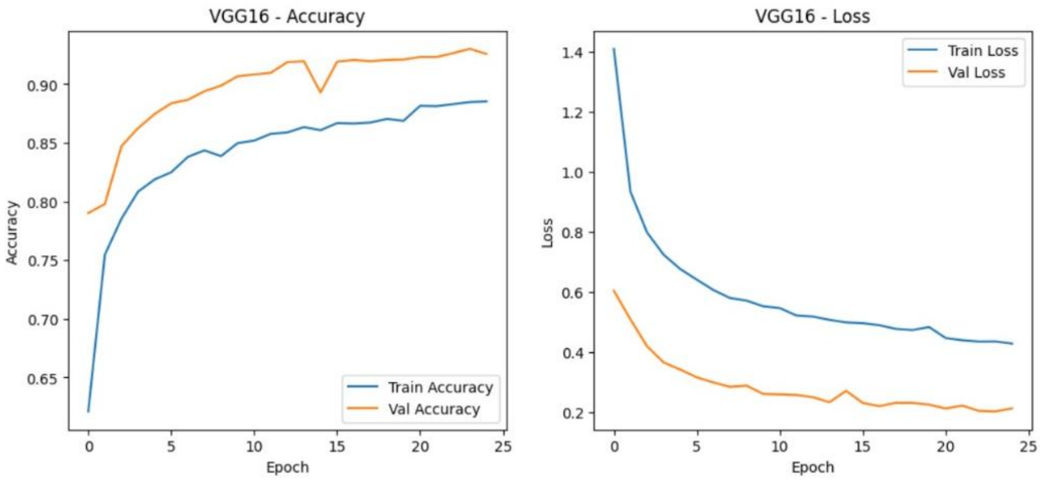

Figure 4.3 Training and Validation Accuracy and Loss graphs of VGG16

4.3.3 RESNET50

The confusion matrix for the ResNet50 model shows how it worked out for classifying into the four categories of chest diseases: "COVID," "Normal," "Pneumonia," and "Tuberculosis." A classification breakdown for it reads as follows:

- COVID Class: Of the samples in the COVID class, the model correctly classified 400 instances as COVID. It misclassified 260 instances as "Normal," and it mislabeled 60 instances as "Pneumonia." This suggests a moderate overlap between the COVID and Normal classes, possibly because of common features.
- Normal Class: In the Normal class, ResNet50 performed satisfactorily by classifying 560 samples correctly. However, an enormous number of samples went wrong since 150 samples were classified as COVID, and 300 as Pneumonia. That proves that in some instances, the model fails to identify Normal cases as well as Pneumonia cases.
- Pneumonia Class: It classified 830 as Pneumonia rightly, but mistakenly classified 10 as COVID and 16 as Normal. It indicates that it has great discriminative capability for the class Pneumonia in comparison to others.
- Tuberculosis Class: The model had an extremely hard time with the Tuberculosis class. Among all the Tuberculosis samples, 48 were classified as COVID, 76 as Normal, and 16 as Pneumonia, and none as Tuberculosis. This is a significant weakness in identifying this class.

Generally, ResNet50 is of medium capacity for classification and failed with the class "Tuberculosis" and overlapped results between classes "COVID" and "Normal." In general, these outcomes show this model has numerous disadvantages in capturing fine detail to attain perfect classification over all categories.

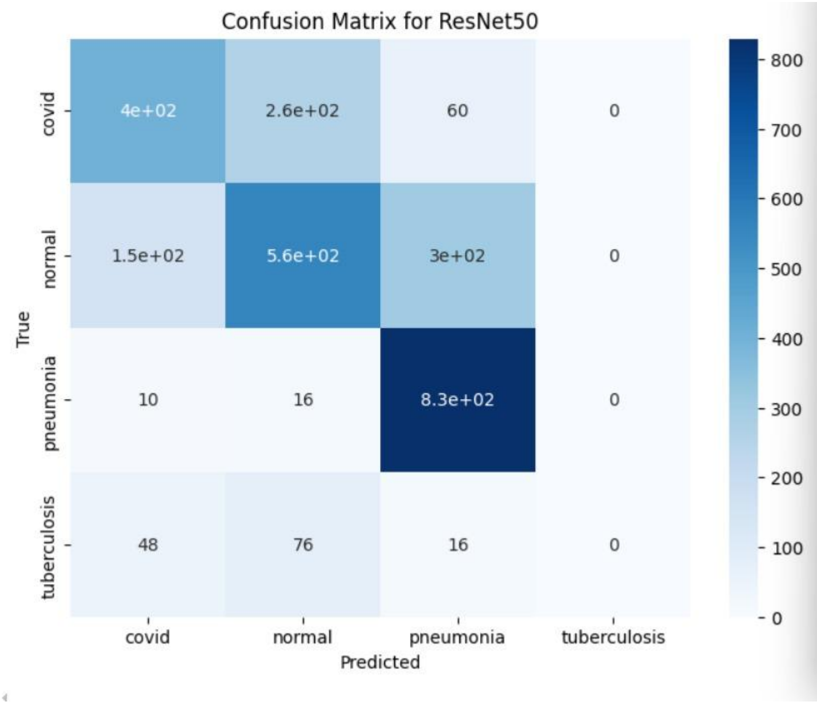

Figure 4.4 Confusion Matrix of ResNet50

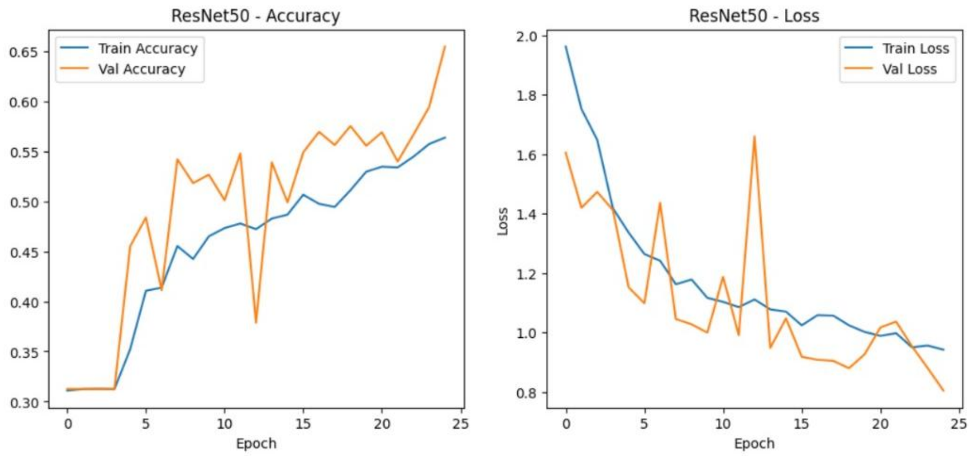

Figure 4.5: Training and Validation Accuracy and Loss graphs of ResNet50

4.3.4 EFFICIENTNETB0

The confusion matrix of the model EfficientNetB0 represents how accurately it classified all the four types that show chest diseases: "COVID," "Normal," "Pneumonia," and "Tuberculosis." In the following are detailed descriptions of them.

- COVID Class: It was unable to classify a single instance as COVID. In the case of samples of COVID, 720 misclassifications took place that were

"Pneumonia." This indicates that the model is unable to make the difference between the class COVID completely.

- Normal Class: The model further classified all the Normal samples as abnormal. All 1,000 Normal instances were classified as "Pneumonia." This clearly indicates that the model totally lacked the capability to distinguish the Normal class.
- Pneumonia Class: All instances were correctly classified as "Pneumonia" by EfficientNetB0, but this also included its actual class, Pneumonia. So even though 860 instances were correctly classified as Pneumonia, failure to distinguish other categories implies drastic overfitting or a basic problem in feature extraction.
- Tuberculosis Class: As with all others, all samples of Tuberculosis were misclassified as "Pneumonia" and 140 instances were incorrectly classified.

It can easily be seen from the confusion matrix that EfficientNetB0 has mistaken all instances of other classes into the Pneumonia category. Overall, the performance of the model is bad and there's no category separability. This might be because the model has not learned meaningful features in training or it faces an extreme class imbalance or something else wrong in the dataset.

This behavior of EfficientNetB0 makes it unsuitable for multi-class classification in this project and requires either considerable retraining or reconfiguration or reconsideration of its suitability for this task.

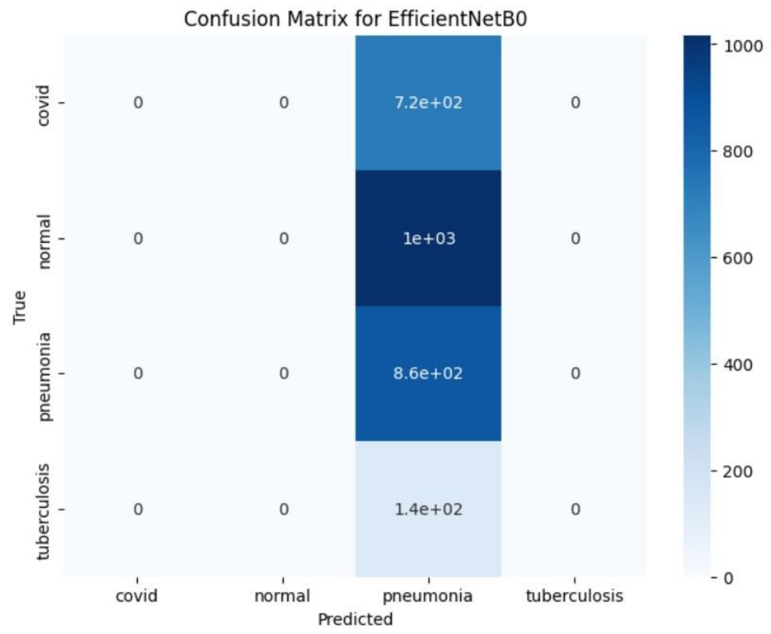

Figure 4.6 Confusion Matrix of EfficientNetB0

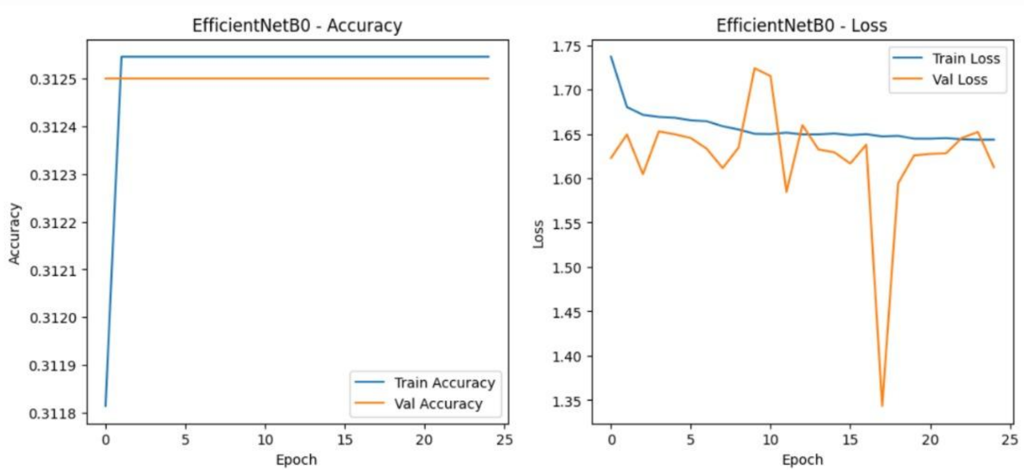

Figure 4.7 Training and Validation Accuracy and Loss graphs of EfficientNetB0

4.3.5 CUSTOM CNN

The confusion matrix of Custom CNN indicates how well the model could classify four types of chest diseases: COVID, Normal, Pneumonia, and Tuberculosis. That table is given below:

- COVID Class: There are no correct classification examples into the class "COVID". Instead, the model misclassifies 660 cases as "Normal," 60 cases as "Pneumonia," and 9 cases as "Tuberculosis". Thus, it indicates the model completely fails in finding the class "COVID".
- For the Normal class, the model performed fairly well and classified 980 samples correctly. It misclassified 1 sample as "COVID," 18 samples as "Pneumonia," and 20 samples as "Tuberculosis."
- Pneumonia Class: The model classified 600 Pneumonia samples correctly but misclassified 240 instances as "Normal" and 10 instances as "Tuberculosis." The high number of misclassifications as Normal suggests a degree of feature similarity between these two classes.
- Tuberculosis Class: For Tuberculosis, the model got 64 correct. However, it misclassified 64 instances as "Normal," 12 as "Pneumonia," and none as "COVID." Therefore, it indicates that the model has a moderate challenge in distinguishing Tuberculosis from other classes, particularly Normal and Pneumonia.

The Custom CNN model is doing an excellent job on the Normal class. Huge classification failure in the COVID class; all instances are misclassified.

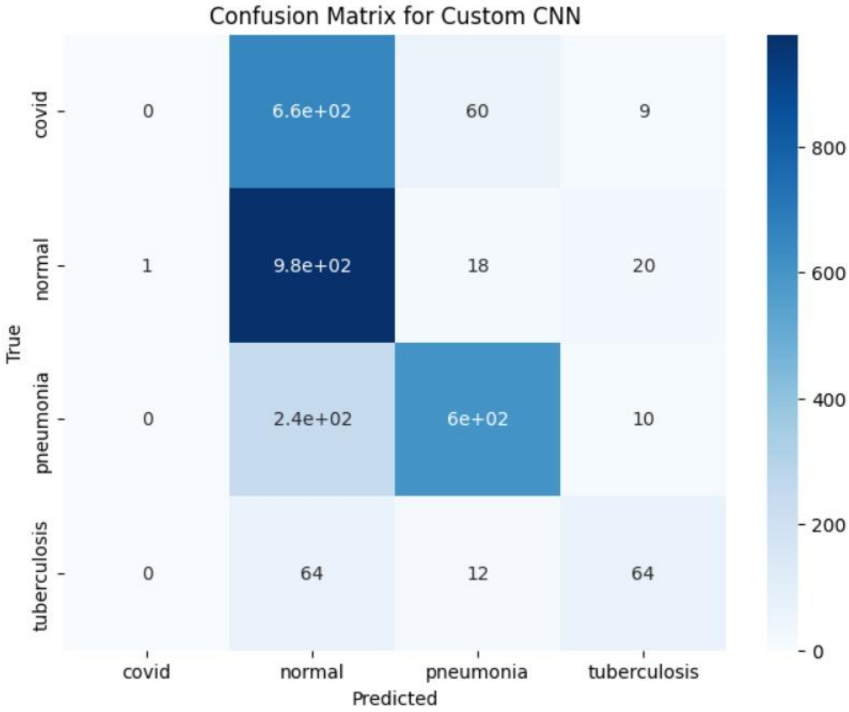

Figure 4.8 Confusion Matrix of Custom CNN

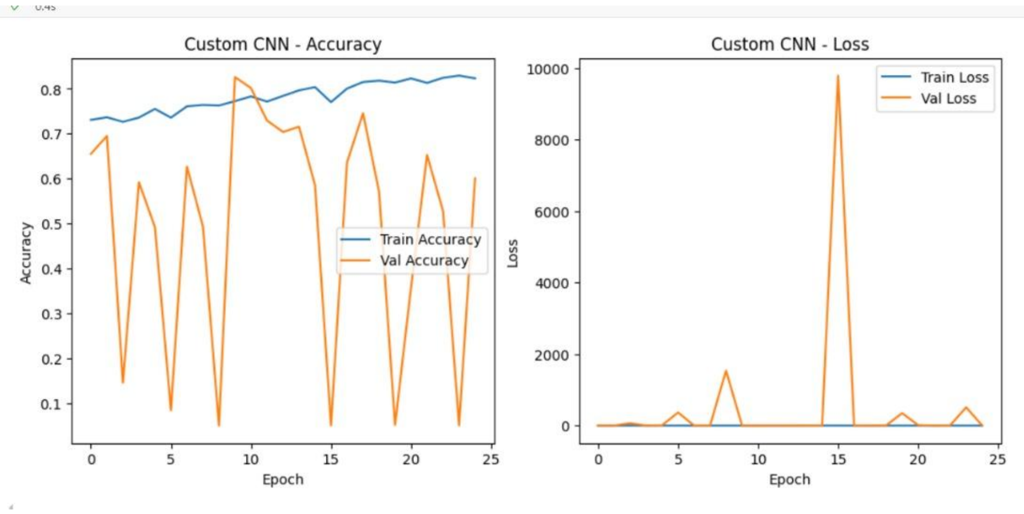

Figure 4.9 Training and Validation Accuracy and Loss graphs of Custom CNN

4.3.6 COMPARATISON OF MODELS ACCURACY

Accuracy comparison is one of the keys in the evaluation of a variety of deep learning models pertaining to the generation of automated reports for chest X-rays. The four deep learning models used in this study included VGG16, ResNet50, EfficientNetB0, and Custom CNN. Their accuracy computed in percentage terms of correct prediction was used to determine their adequacy for real-world use in the classification of diseases of the chest-COVID, Normal, Pneumonia, and Tuberculosis.

A comparison of the accuracies between the models is demonstrated below in Figure 4.10.

Models	Accuracy
VGG16	92.3 %
ResNet50	65.4%
EfficientNetB0	31.2%
Custom CNN	60.0%

Figure 4.10 Accuracy of Models

VGG16:

The highest accuracy achieved was 92.3% by VGG16, and hence it is the most reliable model for this research. Its enhanced feature extraction capability made it classify all four categories of diseases accurately, and hence it proved to be robust and precise. This high accuracy portrays that VGG16 will be a good model for such applications in medical diagnostics which require accurate classification.

ResNet50:

ResNet50 reached an accuracy of 65.4% and moderately well across classes, but the model was confused between categories like "COVID" and "Pneumonia." Though it lacks some capabilities, ResNet50 is still a viable candidate for applications in which there is a trade-off between accuracy and computational efficiency.

EfficientNetB0:

EfficientNetB0 was very poor, achieving accuracy of only 31.2%. It had a lot of struggle with the classification problem; this is well illustrated in the confusion matrix. Such low accuracy means that the network did not learn any characteristic features for the chest diseases and thus is not good to be deployed without serious retraining or optimization.

Custom CNN:

The accuracy of the customized CNN was 60.0%. It outdid the other models, which were EfficientNetB0, but still lags behind ResNet50 and VGG16. For both "Normal" and "Pneumonia," it performed decently. However, for "COVID" and "Tuberculosis," it didn't do well in classifying them. This shows hope for better fine-tuning and architectural adjustment in that particular model.

CONCLUSION FROM THE ACCURACY COMPARISON

This should be clear from the fact that VGG16 performed the best in this context, with an accuracy score of 92.3%. Indeed, as per the above results, although ResNet50 and Custom CNN performed moderately, it is not good enough in terms of accuracy for an application like medical diagnosis with high stakes. EfficientNetB0 performed the poorest and requires significant improvement for its deployment.

4.3.7 DISCUSSION OF FINDINGS

Based on the findings, deep models that determine and classify chest diseases--COVID, Pneumonia, Tuberculosis, and Normal-with chest X-ray images performed well. In terms of accuracy, the best score was obtained with VGG16 which had an accuracy of about 92.3%; hence, very appropriate where precision is involved. Good architecture enabled better feature extraction and proper classification.

ResNet50, with an accuracy of 65.4%, showed moderate performance, while Custom CNN (accuracy: 60.0%) performed slightly better than EfficientNetB0 (accuracy: 31.2%). These models struggled to distinguish between disease categories, making them less suitable for high-precision applications.

The results indicate a trade-off between accuracy and usability. VGG16 is the most reliable for deployment, especially in clinical settings that require high accuracy. The study also indicates that visually similar diseases such as Tuberculosis and Pneumonia are hard to distinguish, thus offering opportunities for future research to improve classification models.

In summary, VGG16 is the most suitable model for this project, offering the highest accuracy and practicality for automated chest X-ray analysis.

4.4 RESULTS OF CHEST DISEASE DETECTION IN THE WEB APPLICATION

The website will thus serve as an easy way through which users upload their chest X-ray images, thereby the automatic diagnosis and report generation. Following uploading of the image, the system is to automatically process the image with the given prediction about the disease, along with its confidence score. The process then helps the users come up with a rapid assessment about their health condition as represented in the X-ray.

In addition to this report of prediction, a full medical report is produced. A report generates the primary findings; probably some implications of the condition detected and a recommendation for further follow-up actions; it helps users attain better knowledge about the diagnosis of a disease and the next move necessary.

Apart from the diagnosis, users can download it in PDF format to refer or share with health workers for reference. The platform is so designed that both visual and textual feedback can be availed upon while enabling decisions based on health.

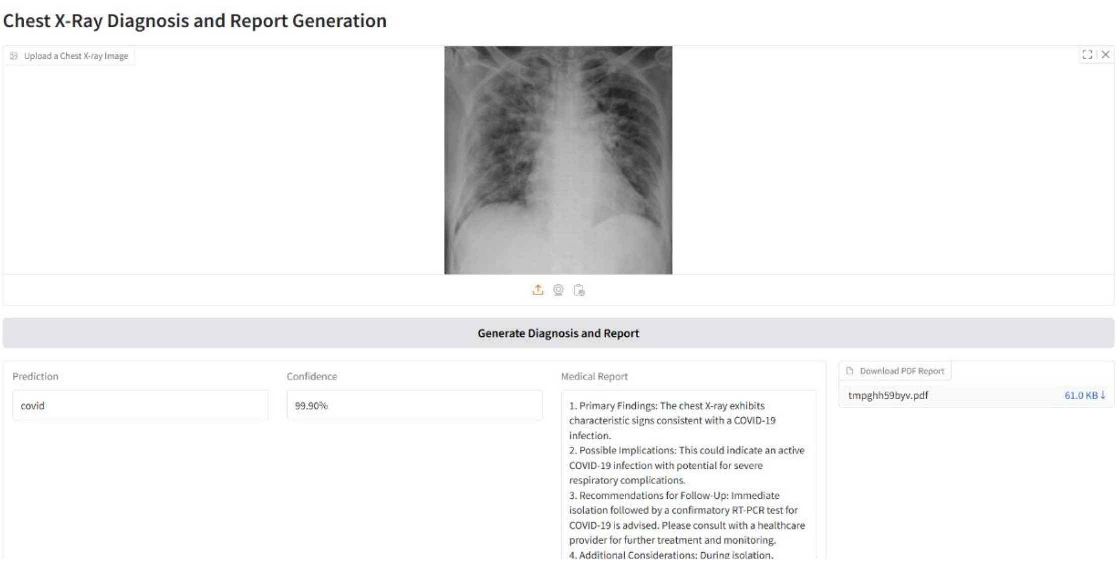

Figure 4.11 Functionality of the App

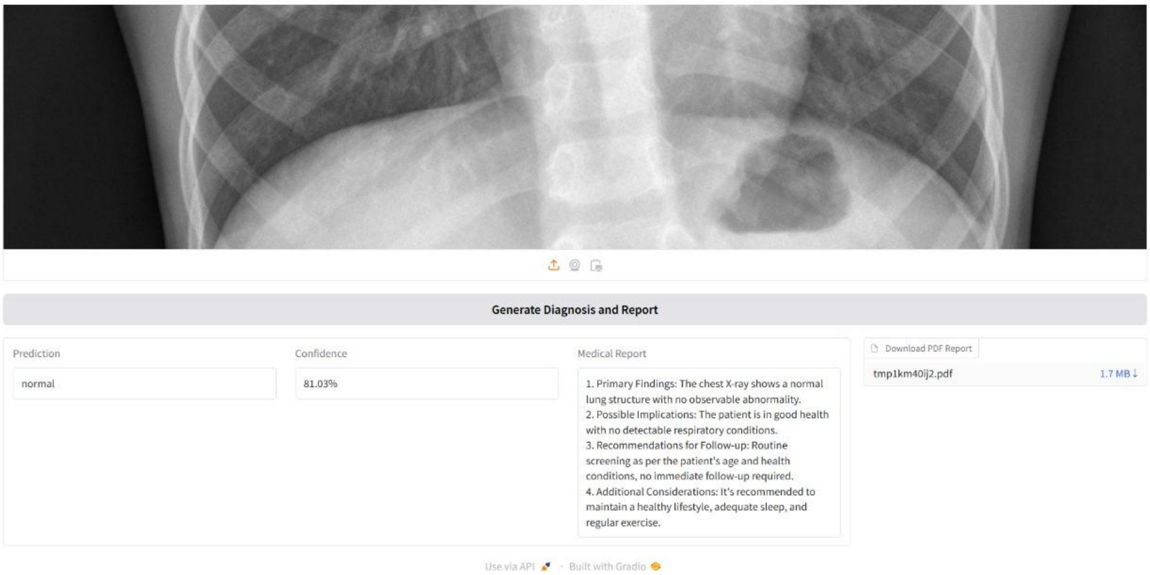

Figure 4.12 Detected Disease

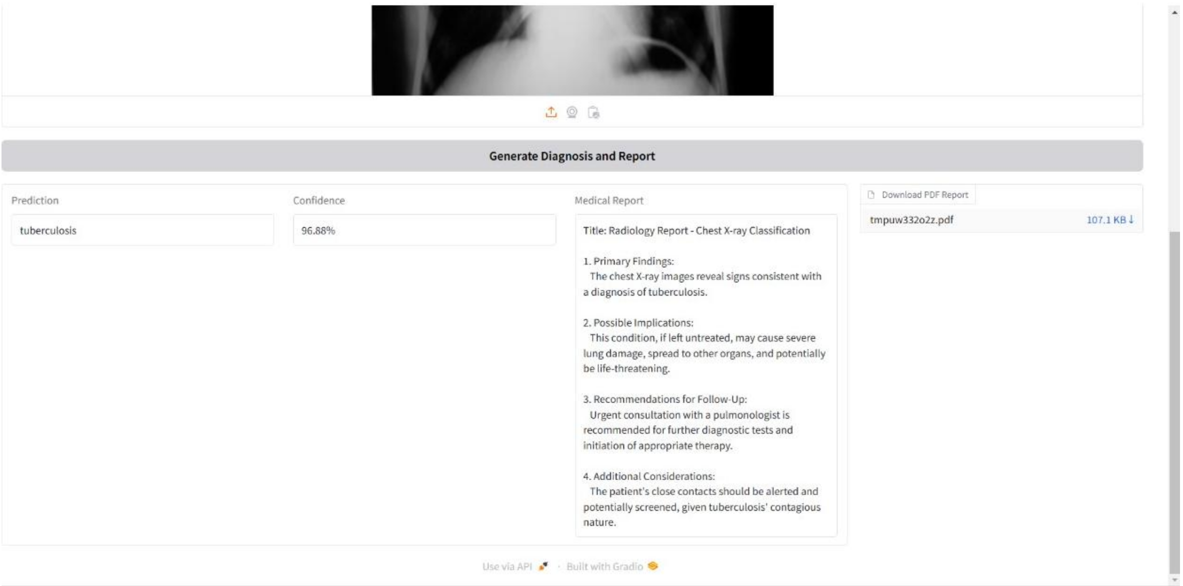

Figure 4.13 Generated Report

CHAPTER 5 CONCLUSION AND FUTURE RECOMMENDATIONS

5.1 CONCLUSION

This project was, hence, able to be an exploratory one wherein integration of deep learning methods would be carried out by this project to generate completely automated radiology reports from a chest X-ray. Deep Learning Techniques: The usage of advanced machine learning model capabilities such as CNN, EfficientNet, ResNet and VGG to help further improve the detection of these diseases like pneumonia, TB and COVID-19 over X-rays of chest. This project thus demonstrates that AI can aid healthcare professionals in diagnosing and reporting medical conditions much more accurately and efficiently and with much less time needed for diagnosis, thereby ultimately improving the care of the patient. It is with the help of systematic training of four different deep learning models that a robust framework for disease classification was developed. The models are trained on an exhaustively chosen dataset of chest X-rays and verified using critical classification metrics. It primarily sought to improve and test the model toward achieving accuracy and generalizability when using it in real-world health-care practices.

In addition, this project also encompasses using GPT-4 to produce reports automatically in radiology, which again describes possibilities of applications of AI toward streamlining the workflow process of diagnosis. Such automation of the generation of complete and contextually precise reports supports the healthcare providers to make fast decisions which can further improve productivity along with reducing the cognitive loads on clinicians.

A user-friendly web-based interface, developed using Gradio, provides an intuitive platform for users to upload chest X-ray images, select a model, and receive disease predictions along with downloadable reports. This user-centric approach ensures that technology is both accessible and feasible for use in real world applications, especially in low resource environments where the experience of radiologists may not be available.

From this, it is demonstrated how machine learning and natural language processing can be combined in order to automate and optimise medical diagnostics. That makes this research project among others a contributing factor for growing knowledge at the interfaces between AI and healthcare. The approach of this study presents a wide open-doors opportunity for future investigations aimed at increasing the accuracies and efficiency of a diagnostic system. Future work potential areas include improving the model performance, transfer learning in case of limited data, as well as the extension of the system to other diseases or modalities. Conclusion This project laid a solid foundation for future uses of AI in health diagnostics. The tool of disease detection and radiology report generation will be extremely useful in this field. As healthcare continues to reap the benefits of the advances made in AI, projects such as this one will become critical in making health care delivery faster, more accurate, and accessible.

The journey to the complete optimization of AI-driven diagnostic systems has just begun, and research will unlock new possibilities in the medical field.

5.2 FUTURE RECOMMENDATIONS

Future work in this project can be directed towards the following areas for enhancing its performance and functionality of the automated radiology report generation system. The first is the ability to include additional information, like the variety of chest X-ray images taken from various sources or from different environmental conditions, to enhance the generalization ability and accuracy in the disease detection process. Also useful will be metadata such as patient demographics or clinical history to give the context, which might refine predictions and enhance diagnostic accuracy. Much room for improvement exists for optimizing the deep learning models used here. Fine-tuning these CNN, EfficientNet, ResNet, and VGG models further may offer improved performance by exploring the possibility of alternative architectures or hybrid models that combine strengths from multiple frameworks. Experiment with newer versions or customized versions of such models to assure real-time performance with better detection accuracy.

Another area that could be improved is creating a mobile application that is a companion to the web version. Such an application may enable uploading chest X-rays from a mobile device itself, thereby allowing for even faster analysis and feedback to healthcare professionals. Real-time notifications regarding detected diseases might also assist healthcare professionals to take appropriate action in a timelier manner, thus caring for the patient better with less risk of misdiagnosis.

This would expand the system in scope and utility for more types of medical imaging applications, such as CT scans and MRI images. Even more versatile and easily applied in a wide number of healthcare settings, would be if frameworks for diseases could be developed, which would be generalizable enough to be detected under any of these imaging modalities.

Field testing and validation of the system through collaboration with healthcare professionals and institutions would be very helpful. Feedback from real-world users involving clinical experts and radiologists would ensure that the system fits clinical needs, works properly in healthcare workflow systems, and can help better fashion the user interface and experience of the system in order to make the tool more practical, reliable, and easy to apply in a real-world clinical medical settings.

Expansion and improvement both in terms of technical capabilities and in real-world applicability lie great in this system. The future of automated radiology report generation promises to be exciting for improving healthcare efficiency, diagnostics, and patient outcomes.

REFERENCES

[1] World Health Organization, “Tuberculosis Key Facts,” WHO, 2022. https://www.who.int/news-room/fact-sheets/detail/tuberculosis

[2] Global Tuberculosis Report 2021, WHO. https://www.who.int/tb/publications/global_report

[3] Centers for Disease Control and Prevention, “COVID-19 Overview,” CDC, 2021. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview

[4] W. R. Webb et al., *Fundamentals of Body CT*, Elsevier, 2021.

[5] M. Lundervold and E. Lundervold, “An overview of deep learning in medical imaging,” *Frontiers in Oncology*, 2019.

[6] Radiological Society of North America, “Chest Radiography,” RSNA. https://www.rsna.org/en/encyclopedia/encyclopedia-of-diseases-and-conditions/chest-radiography

[7] S. Shrimali et al., “Deep learning techniques for pneumonia detection using chest X-rays,” *Journal of Medical Imaging*, 2022.

[8] WHO, “Tuberculosis and COVID-19,” WHO Report 2022. https://www.who.int/news-room/fact-sheets/detail/tuberculosis-and-covid-19

[9] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” arXiv, 2020.

[10] B. Esteva et al., “Deep learning-enabled automated reporting for medical imaging,” *Nature Medicine*, 2019.

[11] P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv preprint arXiv:1711.05225*, Nov. 2017. https://arxiv.org/abs/1711.05225

[12] X. Liu et al., "Detecting COVID-19 from Chest X-ray images with Deep Learning," *IEEE Access*, vol. 8, pp. 149835–149845, 2020. https://ieeexplore.ieee.org/document/9142487

[13] L. Liu et al., "Deep Learning for COVID-19 Detection in Chest X-rays: A Comprehensive Review," *IEEE Access*, vol. 9, pp. 95612–95628, 2021. https://ieeexplore.ieee.org/document/9463662

[14] A. Tuli et al., "COVID-19 Detection using EfficientNet in Chest X-ray Images," *International Journal of Imaging Systems and Technology*, vol. 30, no. 4, pp. 818–827, 2020. https://onlinelibrary.wiley.com/doi/abs/10.1002/ima.22445

[15] J. Vasquez et al., "Detection of Pneumonia using VGG-16 on Chest X-ray Images," *Biology of Medicine*, vol. 53, no. 1, pp. 1–9, 2019. DOI: 10.1016/j.biomed.2019.03.004

[16] M. Chouhan et al., "Transfer Learning for Pneumonia Detection in Chest X-rays," *Journal of Computational Science*, vol. 39, pp. 123–134, 2020. https://www.sciencedirect.com/science/article/pii/S1877750319317344

[17] Z. Yang et al., "AI-Based Radiology Report Generation using GPT-3," *J. Med. Imaging*, vol. 8, no. 4, pp. 148–155, 2021. https://pubmed.ncbi.nlm.nih.gov/33918748/

[18] P. Shah et al., "Data Augmentation Techniques for Improving the Generalization of Medical Imaging Models," *Journal of Machine Learning in Medicine*, vol. 1, pp. 101–112, 2020. https://www.jmlr.org/papers/volume21/20-050/20-050.pdf

[19] S. Ribeiro et al., "Explainable AI in Healthcare: Applications, Challenges, and Future Directions," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 4, pp. 1485–1498, 2021. https://ieeexplore.ieee.org/document/9204473

[20] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," Proceedings of the IEEE, vol. 109, no. 1, pp. 43–76, 2021. doi: 10.1109/JPROC.2020.3004555.

[21] N. M. Elshennawy and D. M. Ibrahim, "Deep-pneumonia framework using deep learning models based on chest X-ray images," Diagnostics, vol. 10, p. 649, 2020. [Online]. Available: https://doi.org/10.3390/diagnostics10090649.

[22] J.-H. Kim, E. Kim, I.-Y. Kim, E.-J. Choi, and S.-H. Byun, “Changes in the Bronchial Cuff Pressure of Left-Sided Double-Lumen Endotracheal Tube by Lateral

Positioning: A Prospective Observational Study," *Journal of Clinical Medicine*, vol. 10, no. 8, p. 1590, Apr. 2021, doi: 10.3390/jcm10081590.