Software Requirements Specification

On the topic "**Chest X-ray (Covid-19 & Pneumonia)**"

Subject: Advanced Machine Learning

Sr.No	Name	Gr No	Roll No
1	Smit Ramteke	22110313	321055
2	Sumedha Sagbhor	22110576	322057
3	Nikita Sanap	22110882	322061
4	Saif Ustad	22110656	322064

Guided By: Prof. Priyanka More

Table of Contents

1. Introduction

1.1 Purpose

The primary aim of this document is to provide a detailed guide outlining the essential requirements for effectively managing, utilizing, and comprehending the Chest X-ray (COVID-19 & Pneumonia) dataset. By delineating the dataset's functionalities, constraints, and dependencies, this document seeks to streamline its integration into research and diagnostic endeavors pertaining to the detection of COVID-19 and pneumonia. Through clear and comprehensive documentation, the goal is to facilitate efficient utilization of the dataset's resources, thereby contributing to advancements in medical imaging analysis and disease diagnosis.

1.2 Scope

The scope of this document encompasses the Chest X-ray (COVID-19 & Pneumonia) dataset, which comprises a meticulously categorized collection of chest X-ray images representing three distinct classes: COVID-19, Pneumonia, and Normal. It delineates the necessary requisites for accessing, manipulating, and administering the dataset for research and diagnostic purposes. Furthermore, this document provides guidance on leveraging the dataset's resources effectively while ensuring adherence to ethical and regulatory standards governing data usage in medical research.

1.3 Definitions, Acronyms, and Abbreviations

This section elucidates key terms, acronyms, and abbreviations pertinent to the document's content:

COVID-19: Coronavirus Disease 2019, caused by the SARS-CoV-2 virus.

SRS: Software Requirements Specification, a document outlining the requirements for a software system.

RT-PCR: Reverse Transcription Polymerase Chain Reaction, a laboratory technique used for detecting specific RNA sequences.

1.4 References

Acknowledging the sources and authors of the dataset is paramount to recognizing the collaborative efforts and contributions that have facilitated its creation and dissemination. Proper attribution fosters transparency and credibility within the

scientific community, ensuring due recognition of individuals and organizations involved in generating the dataset.

1.5 Overview

This document serves as a comprehensive compendium elucidating both the functional and non-functional requirements of the Chest X-ray dataset. It aims to provide stakeholders with a detailed understanding of the dataset's capabilities, limitations, and dependencies, thereby facilitating informed decision-making and effective utilization of the dataset for research and diagnostic purposes.

2. Overall Description

2.1 Product Perspective

The Chest X-ray (COVID-19 & Pneumonia) dataset serves as an invaluable asset for a diverse range of stakeholders, including researchers, clinicians, and developers. It facilitates the development and refinement of AI-based tools aimed at the early detection and diagnosis of COVID-19 and pneumonia from chest X-ray images. By providing a curated collection of annotated images, the dataset enables the training and validation of machine learning algorithms, thereby advancing the capabilities of medical imaging technology in disease detection and diagnosis.

2.2 Product Features

Three classes of X-ray images: The dataset comprises chest X-ray images categorized into three distinct classes: COVID-19, Pneumonia, and Normal. This classification enables researchers to differentiate between various pathological conditions and healthy states.

Organized into train and test sets: The dataset is structured into separate training and testing sets, facilitating the development and evaluation of machine learning models. This partitioning ensures that the performance of models can be assessed accurately on unseen data.

Accessible for research purposes: The dataset is made readily available to the research community, fostering collaboration and knowledge exchange. Researchers, data

scientists, clinicians, and developers can leverage the dataset to conduct studies, validate hypotheses, and develop innovative solutions for medical imaging analysis.

2.3 User Classes and Characteristics

Users of the dataset encompass a diverse range of professionals with expertise in medical imaging and artificial intelligence. These include:

Researchers: Engaged in exploring novel methodologies and techniques for disease detection and diagnosis using medical imaging data.

Data Scientists: Skilled in analyzing and interpreting large datasets to extract meaningful insights and develop predictive models.

Clinicians: Utilize the dataset to enhance diagnostic accuracy and inform clinical decision-making processes.

Developers: Create AI-based tools and applications for automated image analysis and disease detection.

2.4 Operating Environment

The dataset can be accessed and utilized on a variety of platforms compatible with common programming languages and machine learning frameworks. This includes environments such as:

Python: Widely used for data analysis, machine learning, and scientific computing.

R: Popular for statistical analysis and visualization of biomedical data.

TensorFlow, PyTorch, scikit-learn: Commonly employed machine learning frameworks for building and deploying models.

2.5 Design and Implementation Constraints

Images collected from publicly available resources: The dataset is sourced from publicly accessible repositories and databases, necessitating careful consideration of data quality and integrity.

Limited annotations available for images: While efforts are made to provide accurate annotations for each image, there may be limitations in terms of completeness and consistency.

Variability in image quality and resolution: Images may exhibit variations in quality, resolution, and imaging protocols, posing challenges for algorithm development and validation.

2.6 Assumptions and Dependencies

Assumption: It is assumed that the dataset accurately represents cases of COVID-19, pneumonia, and normal conditions based on available annotations. However, there may be inherent biases or limitations in the dataset that warrant careful consideration during analysis.

Dependency: The utilization of the dataset relies on the availability of compatible machine learning libraries and frameworks for data analysis and model development. Dependencies include software packages for image processing, feature extraction, and model training.

This elaboration provides a deeper insight into the product perspective, features, user classes, operating environment, constraints, assumptions, and dependencies associated with the Chest X-ray (COVID-19 & Pneumonia) dataset.

3. Specific Requirements

The Specific Requirements section will provide the Use Case Reports specifying the 11
Use Cases that make up this system. The Use Cases in the section contains both internal and external pre and post conditions.

3.1 Use-Case Reports

To enhance the visual representation of our E-commerce website in the Software Requirement Specification (SRS), we have incorporated screenshots of features similar to those found on Amazon's platform. These screenshots serve as illustrative references to provide a clear understanding of the intended functionalities and user interface within our project.

It's important to note that these screenshots are for representational purposes only and are not indicative of a direct association with Amazon. The use of these visuals aims to convey the expected look and feel of our E-commerce website, aligning with industry standards and user expectations.

By including these screenshots, our SRS endeavors to provide stakeholders with a visual reference point, fostering a more comprehensive understanding of the proposed features and design elements within our E-commerce platform.

4. Risk Analysis:

4.1 Data Security Risks

Risk: Unauthorized access to sensitive medical data

Description: There is a risk of unauthorized individuals or entities gaining access to the dataset containing sensitive medical information, leading to potential misuse or breach of patient privacy.

Mitigation: Implement robust authentication mechanisms, such as multi-factor authentication, and access control mechanisms to ensure that only authorized users can access the dataset. Utilize role-based access controls to restrict access to specific subsets of data based on user roles and responsibilities.

Risk: Data breaches leading to privacy violations

Description: Data breaches pose a significant threat to the privacy of individuals whose medical data is included in the dataset. Unauthorized access to or disclosure of this information can result in legal and reputational consequences.

Mitigation: Encrypt sensitive data both at rest and in transit using strong encryption algorithms. Regularly audit access logs to detect and investigate any unauthorized access attempts. Implement data loss prevention (DLP) mechanisms to monitor and prevent the unauthorized transfer of sensitive data.

4.2 Data Quality Risks

Risk: Inaccurate or inconsistent labeling of X-ray images

Description: Inaccurate or inconsistent labeling of X-ray images within the dataset can lead to erroneous conclusions during analysis and diagnosis, potentially impacting patient care.

Mitigation: Implement rigorous quality control measures during the annotation process to ensure accurate and consistent labeling of X-ray images. Utilize multiple annotators for each image and establish consensus among annotators to mitigate labeling discrepancies. Incorporate validation checks to identify and rectify labeling errors during dataset curation.

Risk: Bias in the dataset leading to skewed model performance

Description: Bias in the dataset, such as underrepresentation of certain demographics or medical conditions, can result in biased model predictions and inaccurate diagnoses, particularly for minority populations.

Mitigation: Ensure the dataset comprises diverse and representative samples of X-ray images across demographics, including age, gender, ethnicity, and geographic location. Conduct bias assessments to identify and mitigate any biases present in the dataset. Implement techniques such as data augmentation and oversampling to address data imbalances and enhance model generalization.

4.3 Technical Risks

Risk: System failure or downtime impacting dataset availability

Description: System failures or downtime, such as hardware failures or network outages, can disrupt access to the dataset, leading to delays in diagnosis and research activities.

Mitigation: Implement redundancy and failover mechanisms to ensure high availability of the dataset. Utilize distributed storage systems and load balancers to distribute traffic and mitigate the impact of hardware failures. Monitor system performance and uptime proactively to identify and address potential issues before they affect dataset availability.

Risk: Compatibility issues with AI algorithms or tools

Description: Incompatibility between the dataset and AI algorithms or tools used for analysis and diagnosis can hinder the integration and adoption of these technologies, limiting their effectiveness.

Mitigation: Conduct thorough compatibility testing to verify the compatibility of the dataset with a wide range of AI algorithms and tools commonly used in medical imaging analysis. Provide guidelines and documentation for integrating the dataset with AI platforms, including instructions for data preprocessing, model training, and evaluation. Collaborate with AI researchers and developers to address any compatibility issues and ensure seamless integration.

4. Supporting Information

4.1 Data Collection and Preprocessing

Data Collection Sources and Methods:

The Chest X-ray (COVID-19 & Pneumonia) dataset is compiled from various publicly available sources, including medical repositories, research publications, and healthcare institutions. Data collection methods involve accessing and aggregating chest X-ray images from these sources while adhering to ethical guidelines and data usage policies. Efforts are made to ensure the diversity and representativeness of the dataset by incorporating images from different demographics, geographical regions, and imaging modalities.

Preprocessing Steps:

Preprocessing is a crucial step to standardize and prepare the dataset for analysis. Common preprocessing steps applied to the chest X-ray images may include:

Image Rescaling and Standardization: Resizing images to a consistent resolution and normalizing pixel intensities to a standardized range (e.g., 0 to 1) to mitigate variability in image quality.

Noise Reduction and Enhancement: Applying filters and techniques to reduce noise, enhance image contrast, and improve overall image quality.

Anonymization and Privacy Protection: Removing or anonymizing sensitive patient information from image metadata to ensure compliance with privacy regulations (e.g., Health Insurance Portability and Accountability Act - HIPAA).

Data Augmentation: Generating additional training samples through techniques such as rotation, flipping, and zooming to increase dataset diversity and improve model generalization.

Labeling and Annotation: Ensuring accurate and consistent labeling of images with their corresponding classes (COVID-19, Pneumonia, Normal) for supervised learning tasks.

These preprocessing steps aim to standardize the dataset, enhance image quality, and facilitate meaningful analysis and interpretation of chest X-ray images for research and diagnostic purposes.

4.2 Model Evaluation Metrics

Evaluation Criteria:

The performance of AI models trained on the Chest X-ray (COVID-19 & Pneumonia) dataset is assessed using a variety of evaluation metrics, including but not limited to:

Accuracy: The proportion of correctly classified instances among all instances.

Sensitivity (Recall): The ability of the model to correctly identify positive cases (e.g., COVID-19 or Pneumonia) out of all actual positive cases.

Specificity: The ability of the model to correctly identify negative cases (Normal) out of all actual negative cases.

Precision: The proportion of true positive predictions among all positive predictions made by the model.

F1 Score: The harmonic mean of precision and recall, providing a balanced measure of a model's performance.

Area Under the Receiver Operating Characteristic (ROC) Curve (AUC): A metric that evaluates the trade-off between true positive rate and false positive rate across different classification thresholds.

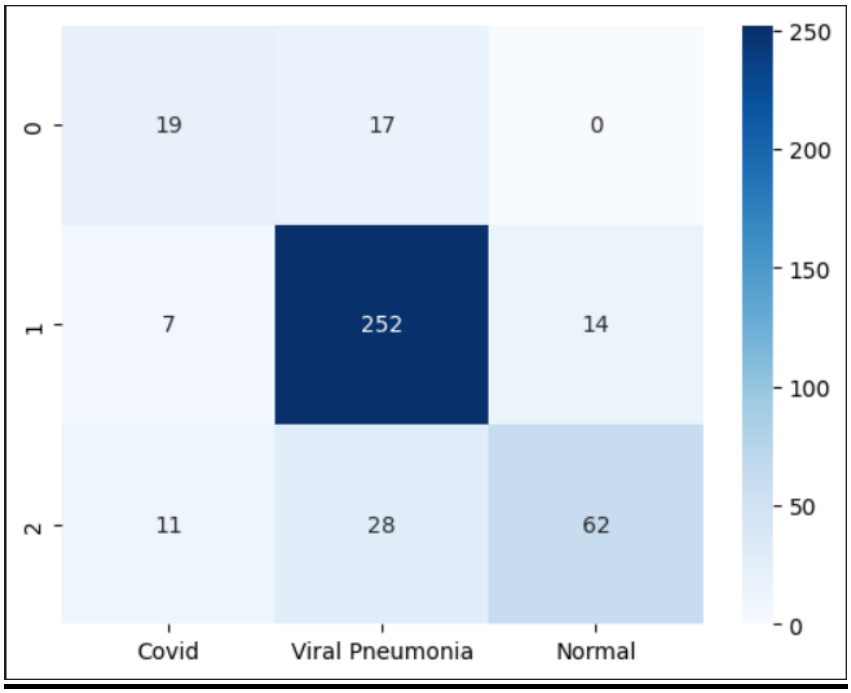
These evaluation metrics collectively provide insights into the performance, robustness, and generalization capabilities of AI models trained on the dataset. They enable researchers and practitioners to quantitatively assess model efficacy and compare different approaches for COVID-19 and pneumonia detection from chest X-ray images.

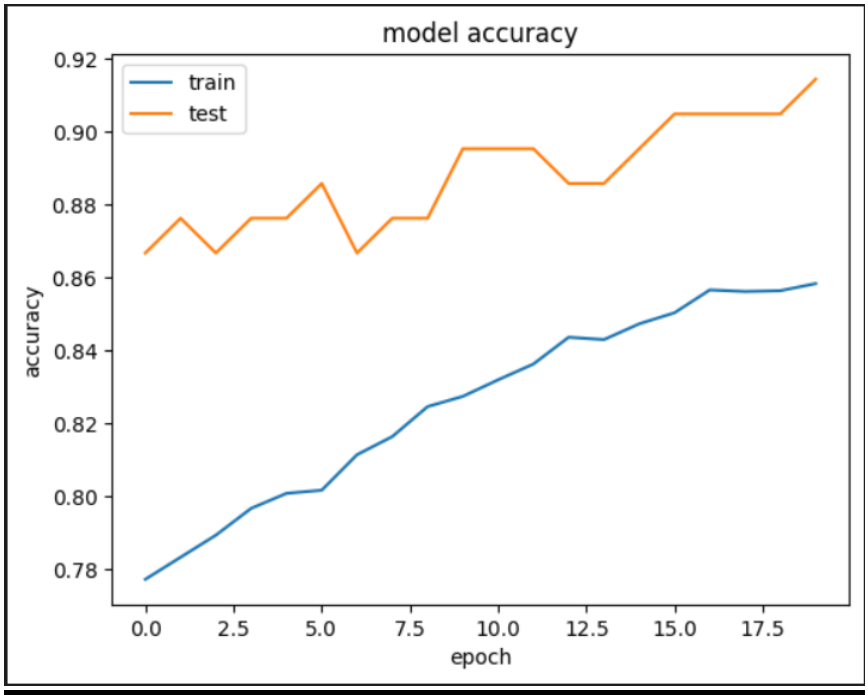
For CNN Model:

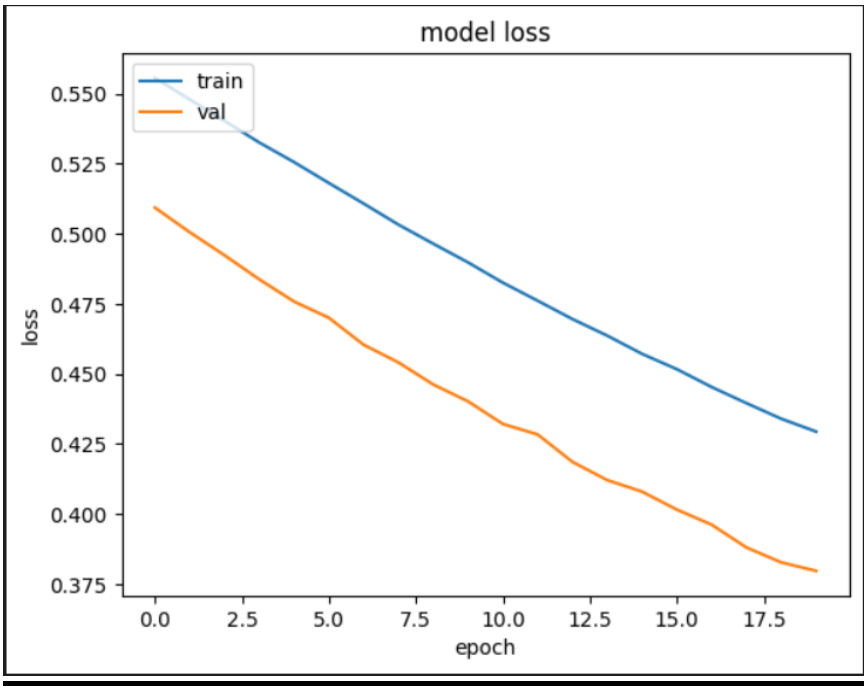
	precision	recall	f1-score	support
Covid	0.94	0.44	0.60	36
Viral Pnemonia	0.85	0.92	0.88	273
Normal	0.82	0.61	0.70	101
Micro avg	0.85	0.80	0.82	410
Macro avg	0.87	0.66	0.73	410
Weighted avg	0.85	0.80	0.81	410
Samples avg	0.80	0.80	0.80	420

Accuracy: 0.8048780487804879

Confusion Matrix:





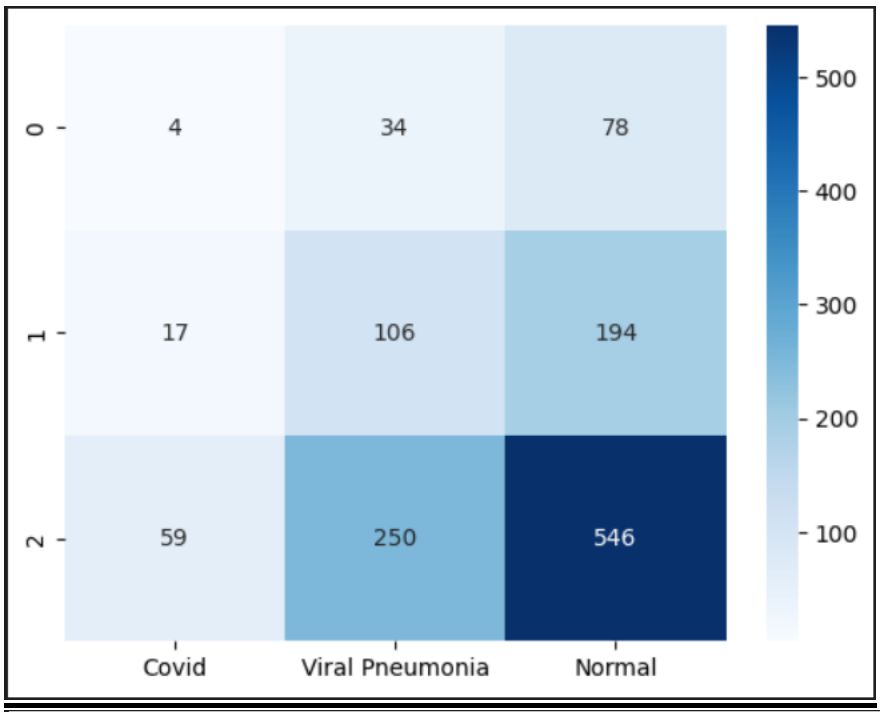


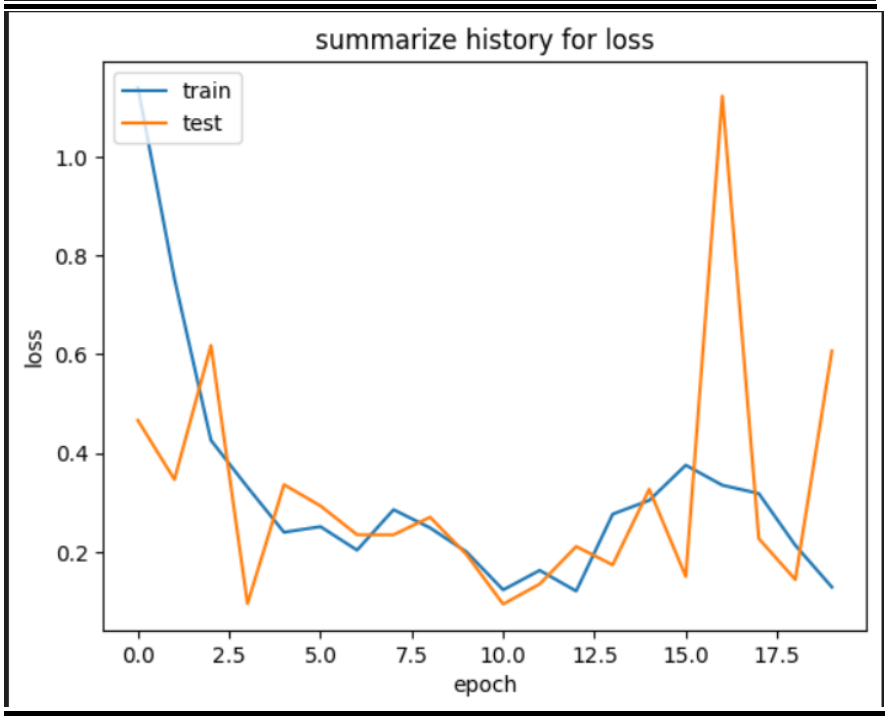
For VGG Model:

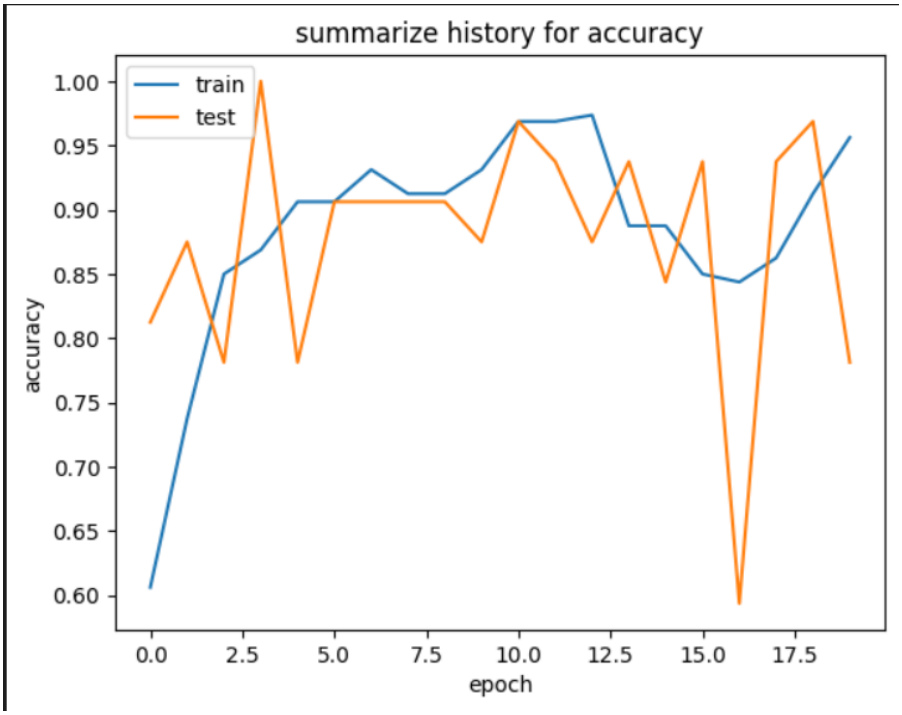
	precision	recall	f1-score	support
0	0.07	0.05	0.06	116
1	0.21	0.26	0.23	317
2	0.64	0.61	0.63	855
Macro avg	0.31	0.31	0.31	1288
Weighted avg	0.49	0.48	0.48	1288

Accuracy : 0.95

Confusion matrix:


0
1
2
4
34
78
17
106
194
59
250
546
500
400
300
200
100
Covid
Viral Pneumonia
Normal


summarize history for loss
train
test
loss
epoch
1.0
0.8
0.6
0.4
0.2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5



VGG 16 vs Generalised CNN :

Architecture:

- **VGG16**: VGG16 is a specific CNN architecture proposed by the Visual Geometry Group at the University of Oxford. It consists of 16 layers, including convolutional layers with small 3x3 filters and max-pooling layers, followed by fully connected layers.
- **Generalized CNN**: A generalized CNN refers to a CNN architecture that may vary in terms of the number of layers, the size of convolutional filters, the use of pooling layers, and other architectural details. It can be customized based on the specific requirements of a project.

Complexity:

- **VGG16**: VGG16 is a relatively deep and complex architecture with a fixed structure, consisting of multiple layers stacked on top of each other. Its fixed structure may limit its adaptability to different datasets or tasks.
- **Generalized CNN**: A generalized CNN offers more flexibility in terms of architecture. It can be adjusted and customized based on the specific requirements of the project, allowing for experimentation with different layer configurations and hyperparameters.

Performance:

- **VGG16**: VGG16 has been widely used and benchmarked on various image classification tasks, including medical image analysis such as chest X-ray classification. It has shown good performance on many datasets.
- **Generalized CNN**: The performance of a generalized CNN can vary depending on its architecture, hyperparameters, and the quality and quantity of training data. With proper design and optimization, a customized CNN can achieve competitive performance on specific tasks.

Resource Requirements:

- **VGG16**: Due to its depth and complexity, VGG16 may require more computational resources (e.g., memory, processing power) for training and inference compared to simpler CNN architectures.
- **Generalized CNN:** Depending on its design and architecture, a generalized CNN can be optimized to balance performance and resource efficiency, making it suitable for deployment on different platforms and environments.

Project Suitability:

- **VGG16**: VGG16 can be a good choice if you prefer a well-established architecture with proven performance on similar tasks and datasets. It is suitable for scenarios where you want a straightforward and reliable solution.
- **Generalized CNN:** A generalized CNN might be preferable if you require more flexibility and control over the architecture and want to tailor the model specifically to your project requirements. It allows for experimentation and fine-tuning to achieve optimal performance.

This elaboration offers a comprehensive understanding of the data collection, preprocessing, and model evaluation aspects associated with the Chest X-ray (COVID-19 & Pneumonia) dataset, emphasizing the importance of standardized procedures and rigorous evaluation criteria for reliable and reproducible research outcomes.

Summary:

The Chest X-ray (COVID-19 & Pneumonia) dataset serves as a critical resource for researchers, clinicians, and developers aiming to develop AI-based tools for early detection of COVID-19 and pneumonia from chest X-ray images. It features three classes of images (COVID-19, Pneumonia, Normal) organized into train and test sets, accessible for research purposes. Users, including researchers, data scientists, clinicians, and developers, can access and utilize the dataset on platforms compatible

with common programming languages and machine learning frameworks. The dataset's design and implementation constraints include image variability and limited annotations, while assumptions center on accurate representation of cases. Supplementary requirements emphasize data quality assurance and documentation, while non-functional requirements focus on performance and security. Supporting information outlines data collection sources and preprocessing steps, along with model evaluation metrics such as accuracy and sensitivity. Overall, the dataset facilitates impactful research and diagnostic endeavors in the field of medical imaging and disease detection.