International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

A Respiratory Disease Management Framework by Combining Large Language Models and Convolutional Neural Networks for Effective Diagnosis

Mohammad Rifat Ahmmad Rashid[1], Mahamudul Hasan[1], Akibul Haque[1], Angon Bhadra Antu[1], Anika Tabassum Tanha[1], Anisur Rahman[1] and M. Saddam Hossain Khan[1]

[1]*Department of Computer Science and Engineering, East West University, Dhaka, 1219, Bangladesh*

Abstract:
Artificial Intelligence in medical diagnostics has the potential to significantly increase patient care and healthcare outcomes. This synergy between advanced artificial intelligence technologies not only optimizes the efficiency of diagnostic analysis but also holds significant promise in improving patient outcomes and supporting healthcare professionals in delivering precise medical interventions. This paper presents a radically new approach that combines the effect of Large Language Models (LLM) with Computer Vision techniques aimed at increasing the performance of medical diagnosis and treatment recommendations for respiratory diseases. For image analysis, we use a pre-trained LLM from Hugging Face, 'Llama-2-7B-chat-GGML', and Convolution Neural Networks (CNN) which consists of InceptionV3, MobileNetV2, and NASNet. The CNN was able to classify chest X-ray images to be 92.85%, 91.88%, and 95.92%. Moreover, the LLM is used to analyze clinical data and generate therapeutic recommendations. We achieved a reduction in inference time of around 33.1% from 165.6 seconds to 111.9 seconds in the most general scenario. Such interaction of CNN and LLM in system use increases the information value of medical diagnostic analysis with high potential for increasing the healthcare outcome. Detailed information about the workflow, diagnostic techniques, and recommendation generation is presented. Experimental analysis of the developed system indicates the application of a combination of LLM and CNN for medical diagnostic purposes to aid healthcare professionals in making informed decisions and providing precise medical advice.

Keywords: Bioinformatics, Natural language processing, Medical diagnosis, Respiratory diseases, Convolutional neural network, Large language models, Patient healthcare

1. INTRODUCTION

The healthcare industry and the field of bioinformatics have experienced significant evolution driven by advancements in artificial intelligence (AI) and natural language processing (NLP) [1], [2], [3]. Despite the promising integration of AI in these areas, the development of a fully realised framework that synergizes the capabilities of large language models (LLMs) with the analytical power of convolutional neural networks (CNNs) for the diagnosis and management of respiratory diseases, such as pneumonia and COVID-19, remains incomplete. The need for such an integrated system is particularly acute in the realm of respiratory illnesses, where rapid and accurate diagnosis is crucial for effective treatment. While several studies have explored the individual benefits of LLMs and CNNs in medical contexts, their combined utility in a cohesive diagnostic and healthcare enhancement framework has not been fully explored or developed. This gap underscores an opportunity for a transformative approach that could substantially improve the outcomes for patients suffering from respiratory conditions by leveraging the strengths of both LLMs and CNNs.

Respiratory diseases, particularly pneumonia and COVID-19, have presented significant challenges to global health due to their prevalence and the critical importance of timely and accurate diagnosis. While effective, traditional diagnostic methods often require considerable time and resources, which can delay the initiation of appropriate treatments. The advent of AI technologies, particularly CNNs, has shown great promise in the rapid and accurate classification of medical images [**?**], [**?**], [4]. Simultaneously, LLMs have demonstrated substantial capabilities in understanding and generating human-like text, making them invaluable in parsing clinical data and providing contextually relevant medical advice [5], [**?**].

In recent years, there has been an increasing focus on the application of CNNs for medical image analysis, with several studies highlighting their efficacy in detecting various conditions from radiographic images [**?**], [6]. For instance, CNN architectures such as DenseNet201 and VGG19 have been employed successfully to classify pneumonia and COVID-19 from chest X-rays with high accuracy [1], [6]. Concurrently, LLMs like GPT-3 have been explored for their potential to enhance clinical decision-making through advanced natural language understanding and generation [2], [3]. However, the integration of these two powerful technologies into a unified framework for respiratory disease diagnosis and management has not been fully realized.

This paper aims to bridge this gap by presenting a novel integrated system that combines the strengths of LLMs and CNNs for the diagnosis and treatment recommendation of respiratory diseases. Our approach leverages the image classification prowess of CNNs with the contextual understanding capabilities of LLMs to provide a comprehensive diagnostic and advisory system. Specifically, we utilize the ’Llama-2-7B-chat-GGML’ model from Hugging Face in conjunction with CNN architectures such as InceptionV3, MobileNetV2, and NASNet to achieve high classification accuracies and generate tailored therapeutic suggestions. The key contributions of this work are:

- Demonstrating the feasibility of integrating LLMs with CNNs for enhanced diagnostic accuracy.
- Providing a detailed workflow of the system, including image analysis, disease classification, and recommendation generation.
- Evaluating the performance of the integrated system in terms of classification accuracy and inference time, showing significant improvements in both areas.

The remainder of this paper is structured as follows: Section 2 reviews the related work in the application of AI in medical diagnostics. Section 3 describes the methodology used in developing the integrated system. Section 4 presents the experimental results and performance analysis. Section 5 discusses the implications of the findings and potential areas for future research. Finally, Section 6 concludes the paper, summarizing the key contributions and impact of the study.

2. RelaTed Work

This section reviews the existing research on AI-driven diagnosis using CNNs and LLMs. It is divided into two subsections: the first is dedicated to the use of CNNs for detecting respiratory diseases, and the second reviews LLMs as tools for generating medical recommendations.

A. CNNs for Respiratory Illness Detection

Convolutional neural networks (CNNs) have demonstrated significant potential in the field of medical image analysis, particularly for the detection of respiratory diseases such as pneumonia and COVID-19. Numerous studies have explored various CNN architectures for their effectiveness in classifying medical images. For instance, DenseNet201 and VGG19 have been widely used due to their deep feature extraction capabilities and robust performance in image classification tasks [1], [**?**]. These architectures have shown high accuracy in distinguishing between normal and diseased states in chest X-ray and CT scan images.

Cheng Wang proposed a method for diagnosing pneumonia by using graph reasoning [7]. By building a graph representation of various lung regions, their proposed model was able to depict the relationships between various lung regions in an X-ray image and accurately detect pneumonia. Alhassan Mabrouk et al. [8] presented an ensemble model in their work that combines the predictions of various CNN architectures. By doing so, they were able to overcome the limitations of any particular CNN model and detect pneumonia with utmost accuracy and lower false positive scores.

Another work done by Salehi et al. [4] proposed an automatic transfer-learning method for pneumonia detection. Among their CNN models, DenseNet121 performed better than other models Xception, VGG19, and ResNet50 for pneumonia classification. Ezaz Khan [6] performed multi-class chest X-ray-based COVID-19 detection by pre-trained deep learning models - EfficientNetB1, NasNetMobile, and MobileNetV2. They also improved each CNN model’s performance by fine-tuning, retraining, and regularization. Their regularized EfficientNetB1 was able to outperform the other models in classification accuracy. Tuan Le Dinh [9] performed COVID-19 classification on a custom dataset using several CNN models - DenseNet, ResNet50, InceptionNet, Swin Transformer, Hybrid EfficientNet-DOLG. They also assessed the accuracy, precision, recall, and F1 scores to validate each CNN architecture’s performance for chest X-ray classification.

The comparative performance of these studies is summarized in Table I, which highlights the accuracy, precision, recall, and F1 scores achieved by each method.

B. LLM-Based Recommendations

Pretrained Large Language Models (LLMs) have emerged as powerful tools for generating human-like texts. GPT-3 [5] achieves strong performance on many NLP datasets, including question-answering, translation, and cloze tests. This model, with its 175 billion parameters, excels in zero-shot, one-shot, and few-shot settings. RecSys-Assistant-Human (RAH) [10], which processes complex user inquiries and generates context-aware interactive recommendations.

LLMs can encode textual features for enhanced usage, item representations, and recommendation purposes. Hao Ding [11] proposed a ZeroShot recommender system that learns user behavioral patterns and generalizes across datasets using BERT. This system provides relatable item recommendations. Yupeng Hou [12] evaluated the promising zero-shot ranking abilities of LLMs by constructing natural language prompts with historical interactions. They found that LLMs could perceive the order of sequential

TABLE I. Comparison Table of Related Classification Methods in Detecting Pneumonia and Covid-19

Reference	Name	Accuracy	Precision	Recall	F1 score
C. Wang [7]	Efficient Graph Model	0.89	0.89	0.91	0.90
A. Mabrouk [8]	Ensemble learning	0.9391	0.9396	0.9299	0.9391
M. Salehi [4]	DenseNet121	0.868	0.868	-	0.898
Ejaz Khan [6]	EfficientNetB1	0.9613	0.9725	-	0.975
T. Le Dinh [9]	Hybrid EfficientNet- DOLG	0.95	0.95	0.96	0.95

interactions and utilize them effectively for recommendation purposes.

Early research in LLM-based recommendations includes M6-Rec, an open-ended language model [13], and P5 [14], which unifies various recommendation tasks into a shared language modeling and NLP framework. Zhang et al. utilized GPT-2 [15] or BERT [16] as the recommender engine for next-movie prediction.

LLMs have shown remarkable potential to significantly enhance medical healthcare facilities and bioinformatics research [17]. LLMs can encode clinical knowledge for better understanding [18] and extract biomedical information from knowledge graphs [19]. BERT, for instance, performed a highly accurate classification of chest radiographic reports by pretraining on 3.8 million text reports [1]. Malik Sallam [2] reviewed the potential use of the GPT-3 model for healthcare education and research. GPT-3 can also perform successive computer-aided diagnoses on medical images [20].

More modern LLMs, including OPT (ranging from 125M to 175B parameters) [5], PaLM (540B parameters) [21], have shown impressive results in diverse NLP tasks such as logical reasoning, problem-solving, and unsupervised multitask learning. Another LLM model, UniSrec [22], learns universal representations of user behavior sequences using pre-trained large language models. Youpeng Hou demonstrated the VQ-Rec framework [23], which obtains text encodings via language models and maps them for embedding lookups. ChatRec [24] employs conversational multi-round recommendations using LLMs, showing that LLMs can enhance cross-domain recommendation systems by improving interactivity and explainability.

The knowledge sphere of modern LLMs has broadened compared to earlier language models. Several studies have already shown preliminary results using LLMs as recommenders with task-specific prompts. Junling Liu [25] assessed the performance of GPT-3 as a recommendation engine compared to traditional recommendation models. Their analysis shows that GPT-3 outperforms traditional methods in intelligible recommendation tasks and its potential to generate explanations. LLMs are also capable of providing personalized recommendations based on user instructions [26]. Damien Sileo [27] proposed a different recommendation framework that only uses unstructured text corpora as training data. The work concludes that standard language models can perform next-item recommendations compared to standard matrix factorization on trained data. The various LLM-based recommendation methods and their key findings are summarized in Table **??**.

C. *Summary of Related Work*

Respiratory disease detection has seen increasing use of convolutional neural networks with studies using architectures like DenseNet201 and VGG19. Moreover, Wang's graph reasoning model for pneumonia and Mabrouk et al.'s ensemble have also examined CNN application in this modality. In addition to these examples, Salehi et al. used transfer learning with DenseNet121, and Khan's multi-class COVID-19 detection with EfficientNetB1 observed increased performance through fine-tuning. Table shows the comparison of classification methods to detect pneumonia and COVID-19 disease.

In the domain of recommendation and generated text, large language models have seen concurrent advances. Models like GPT-3 have been employed for a range of NLP tasks, including question-answering and translation, proving their efficacy in zero-shot and few-shot learning. Other LLM-based frameworks like RAH and InteRecAgent have been designed for contextual question-answering and interactive recommendation, respectively. In the context of medical healthcare, LLMs have been harnessed for clinical knowledge encoding, biomedical information extraction, and even the classification of radiographic reports, as seen with BERT pre-trained on text reports. GPT-3's capabilities extend to aiding healthcare education and successive computer-aided medical diagnoses. Table shows a summary of the various methods along with their findings.

In summary, modern LLMs such as OPT, PaLM, and LlaMA show rapid results across NLP tasks. Their development has seen the expansion of the knowledge sphere, with LLMs providing recommenders for various studies and demonstrating effective use even in task-specific prompts. Liu's evaluation showcased GPT-3's performance over traditional baselines in intelligible tasks, and Sileo's systems show that LLMs compare with standard recommendation methods. This paper builds on the development of Llama-2-

TABLE II. Summary of LLM-Based Recommendation Methods

Ref.	Method	Key Findings	Application
Brown et al. [5]	GPT-3	Strong performance on NLP tasks	Question-answering, Translation
Shu et al. [28]	RecSys-Assistant-Human (RAH)	Effective human-like recommendations	Contextual Question Answering
Huang et al. [10]	InteRecAgent	Context-aware interactive recommendations	Conversational Recommendation
Ding et al. [11]	ZeroShot Recommender	Generalizes across datasets using BERT	Item Recommendations
Hou et al. [12]	Zero-shot Ranking	Perceives historical interactions	Sequential Interaction Ranking
Cui et al. [13]	M6-Rec	Open-ended language model for recommendations	Various Recommendation Tasks
Geng et al. [14]	P5	Unifies recommendation tasks in NLP framework	Next-Item Recommendation
Bressem et al. [1]	BERT	High accuracy in chest radiographic reports classification	Medical Image Analysis
Sallam et al. [2]	GPT-3	Potential in healthcare education and research	Healthcare Education, Research
Touvron et al. [21]	LLaMA	Impressive results in diverse NLP tasks	Logical Reasoning, Problem-Solving
Hou et al. [22]	UniSrec	Learns universal user behavior sequences	Behavioral Sequence Recommendation
Hou et al. [23]	VQ-Rec	Embedding lookups via language models	Cross-domain Recommendation
Gao et al. [24]	ChatRec	Enhances interactivity and explainability	Conversational Multi-Round Recommendations
Liu et al. [25]	GPT-3	Outperforms traditional recommendation models	Intelligible Recommendation Tasks
Zhang et al. [26]	Personalized Recommendations	Provides personalized recommendations based on user instructions	Personalized Recommendations
Sileo et al. [27]	Unstructured Text Corpora Framework	Effective next-item recommendations	Next-Item Recommendation

7B-chat-GGML for use in medical recommendation, using its 7 billion parameters to provide contextually relevant advice based on respiratory illness diagnoses such as pneumonia and COVID-19, aiming to enhance patient care with better health recommendations.

3. Methodology

The methodology is split into two main parts, with each part contributing to a key element of the AI-supported diagnosis system. As a reference, Figure 1 depicts the high-level architecture of our approach.

Pneumonia Detection: The first component focuses on detecting respiratory diseases, specifically due to pneumonia and COVID-19, utilizing Convolutional Neural Network (CNN) models. The CNN architectures selected for this purpose include:

- **Inception V3 [29]:** This model is renowned for its sophisticated architecture, which includes multiple convolutional layers with varied filter sizes. It leverages inception modules that allow the network to capture intricate patterns and details at multiple scales within the imaging data. Inception V3 is particularly efficient in terms of computation and depth, making it well-suited for detailed image classification tasks required in medical diagnostics.
- **MobileNetV2 [30]:** Known for its lightweight architecture and efficiency, MobileNetV2 is particularly

advantageous in scenarios requiring fast and efficient processing. The model utilizes depthwise separable convolutions, significantly reducing the number of parameters and computational load while maintaining high accuracy. The version used in this study was specifically iterated through 200 epochs, indicating a comprehensive training process aimed at refining its diagnostic precision and robustness in identifying respiratory diseases from X-ray and CT images.

- **NASNet (Neural Architecture Search Network)[31]:** NASNet employs a novel approach where the architecture is optimized through an automated search process. This process identifies the best performing network structures tailored to the specific task of medical image classification. NASNet's capability to evolve its architecture through neural architecture search allows it to achieve superior performance and robustness in disease identification. It is particularly useful for adapting to the varying complexities present in different medical imaging datasets.

LLM Based Recommendation System: The second component of the methodology revolves around the recommendation system, which is underpinned by a pre-trained Large Language Model (LLM). This system is structured into three sequential stages:

- **Medical Context Interpretation:** At this initial stage, the LLM processes input data from the disease detection component, which includes diagnostic results from the CNN models. The model interprets the clinical context by understanding the specific health scenario of the patient, such as the severity of the detected condition, patient history, and other relevant clinical details. This step ensures that the recommendations are personalized and contextually relevant to the patient's current medical situation.

- **Knowledge Base Consultation:** After establishing the medical context, the LLM refers to an extensive and continually updated knowledge base. This knowledge base comprises medical literature, clinical guidelines, best practice protocols, and case studies. By accessing this repository, the LLM aligns its recommendations with the latest medical standards and evidence-based practices. This step is crucial for ensuring that the generated recommendations are not only accurate but also reflect the most current and accepted medical knowledge.

- **Recommendation Generation:** In the final stage, the LLM synthesizes the interpreted medical context with the consulted knowledge base to generate specific recommendations. These recommendations may include potential treatment options tailored to the detected condition, suggestions for further diagnostic tests to refine the diagnosis, or referrals to specialists for advanced care. The output is designed to be actionable and practical, providing healthcare professionals with clear guidance on the next steps in patient management.

A. *Detailed Architecture of the AI-Assisted Diagnostic and Recommendation System*

In essence, the methodology outlined in this paper presents an integrated approach where the initial diagnostic intelligence is provided by advanced image-processing CNN models, and the subsequent patient-specific medical recommendations are produced by a sophisticated LLM system. A detailed architecture of the AI-assisted diagnostic and recommendation system is shown in Figure 2, detailing the CNN model classification process for X-ray detection and the subsequent recommendation generation by the LLM model based on the medical context and knowledge base. It can be broken down into two main modules:

- **Disease Detection:** This module involves the CNN Model Classification, where X-ray images are input to detect potential respiratory diseases. The process leverages state-of-the-art CNN models which are not specified in the image but could include Inception V3, MobileNetV2, or NASNet as previously mentioned.

- **LLM Model:** Once the disease is detected and the medical context is established from the X-ray detection module, the process flows into the Large Language Model (LLM) module. This module has several internal steps:
  - **Knowledge Base:** The LLM model first interacts with a Knowledge Base, which likely contains vast amounts of medical data, guidelines, and literature, to ground the subsequent recommendation in solid, evidence-based medical understanding.
  - **Prompt:** With the context and knowledge in place, a prompt is generated for the LLM. This prompt effectively communicates the specific details and nuances of the detected medical condition, guiding the LLM towards generating a relevant and accurate recommendation.
  - **Recommendation:** In response to the prompt, the LLM model, specified here as 'Llama-7B', processes the information and outputs a recommendation tailored to the detected medical condition. This recommendation could range from treatment options to further diagnostic steps or specialist referrals.

B. *Pneumonia Detection*

In this subsection, we elucidate the application of Convolutional Neural Networks (CNNs) for the classification of X-ray images to identify pneumonia and COVID-19. CNNs, specialized machine learning models for image processing and computer vision tasks, have been utilized for their

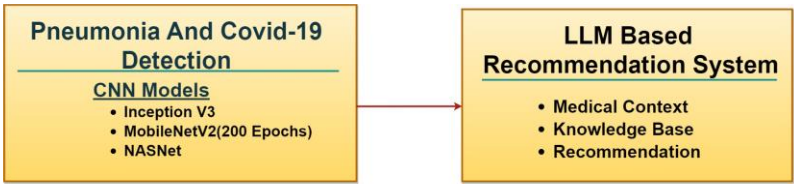
Pneumonia And Covid-19 Detection
CNN Models
• Inception V3
• MobileNetV2(200 Epochs)
• NASNet
LLM Based Recommendation System
• Medical Context
• Knowledge Base
• Recommendation

Figure 1. High Level Architecture.

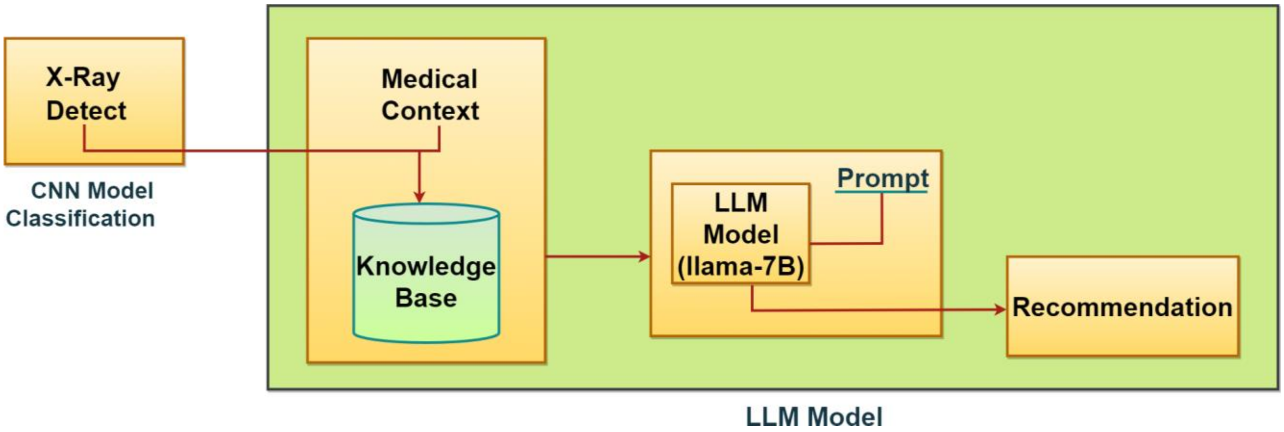

Figure 2. An overview of the AI-Assisted Diagnostic and Recommendation System. The process begins with CNN Model Classification for disease detection from X-ray images, followed by an LLM-based system that utilizes a medical Knowledge Base and prompts to generate tailored recommendations.

proficiency in pattern recognition within complex image data.

The dataset for our CNN model comprises 8,118 chest X-ray images across three categories: normal, pneumonia, and COVID-19, sourced from [32]. The input to our CNN models is a standardized image of size 224x224x3. The CNNs are rigorously trained on this dataset to differentiate among the three aforementioned classes. Figure 3 depicts the CNN models' classification process.

Post-training, the CNN models possess the capability to conduct a precise analysis of X-ray images. The images undergo pre-processing before being input into the CNN models, which include InceptionV3, MobileNetV2 (enhanced through 200 epochs of training), and NASNet. These models execute binary classification, yielding a prediction of either 0 (normal) or 1 (presence of pneumonia or COVID-19). Figure 4 provides examples of the X-ray images used in our study.

C. LLM Recommendation

This subsection outlines the construction of our knowledge base (KB), crucial for diagnosing medical conditions through the analysis of X-ray images and the assimilation of medical context. Our KB encompasses a diverse array of disease symptoms, corresponding treatments, preventive measures, and associated medications. For the KB assembly, data were curated from an array of sources including PDFs, CSV files, and web scraping, drawing from a wide spectrum of medical encyclopedias and authoritative sources such as the World Health Organization's (WHO) medical articles. Footnotes provide the URLs to these valuable resources.[1] [2] [3] [4] [5]

The diagram (Figure **??**) showcases the utilization of a large language model (LLM) to synthesize patient-centric recommendations. Within our system's pipeline, the LLM discerns insights from the KB and, through a sophisticated prompt strategy, formulates recommendations that are customized to the patient's specific health context. These recommendations span medication prescriptions, suggested lifestyle modifications, and potential further diagnostics required.

We emphasize that our LLM of choice is the LLaMA2-7B model, preferred over larger counterparts like GPT-3 due to its favorable balance between size and performance. Its relatively modest scale enhances both the training and deployment efficiency. Opting for LLaMA2-7B with its 7 billion parameters—instead of the more substantial LLaMA2-13B model—allows for expedited and

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7138033/
[2] https://onlinelibrary.wiley.com/doi/book/10.1002/0470114207
[3] https://infobooks.org/free-pdf-books/medical/
[4] https://worldofmedicalsaviours.com/mbbs-pdf-books/
[5] https://www.who.int/publications-detail-redirect/9789241210157

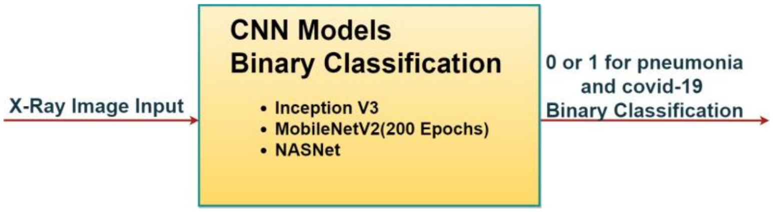

Figure 3. Architectural flow of CNN models for Pneumonia and COVID-19 detection.

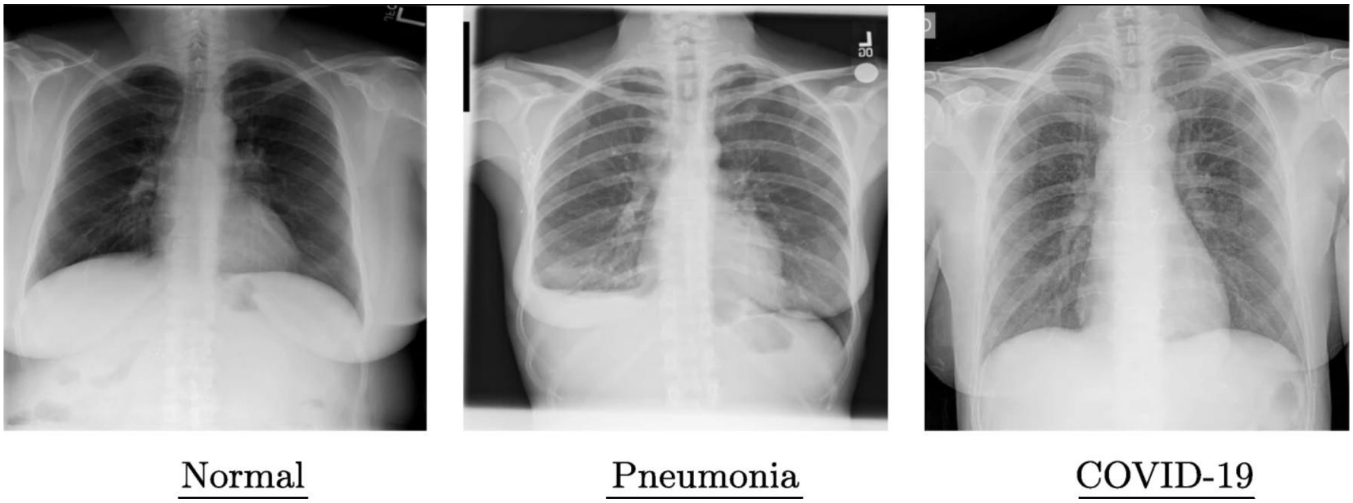

Figure 4. Representative samples of X-ray images used as inputs for the CNN models.

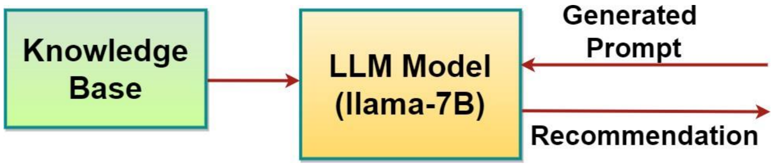

Figure 5. Workflow of the LLM Recommendation Model.

more resource-efficient operations on standard PC CPUs. Our deployment was facilitated by hardware comprising a Ryzen 5 2400G CPU with 16GB RAM, demonstrating the model's practicality for diverse computational environments. By leveraging the LLaMA2-7B model, our system can efficiently generate precise and context-aware recommendations. This includes:

- Medication Prescriptions: Based on the detected condition and referenced medical guidelines.
- Lifestyle Modifications: Suggestions for diet, exercise, and other habits tailored to the patient's needs.
- Further Diagnostics: Recommendations for additional tests or specialist consultations to ensure comprehensive care.

D. Data Preprocessing

Figure 6 illustrates the data preprocessing workflow for a recommendation system powered by a large language model (LLM), specifically 'Llama-7B'. The process begins with the collection of custom data from various formats, including CSV files, PDF documents, and website content. This data, primarily textual in nature, undergoes transformation into a numerical form represented as multi-dimensional vectors, commonly known as embeddings. These embeddings serve as a standardized input for efficient processing and understanding by the LLM.

Subsequently, the embeddings are stored in a vector database, indicated in the diagram as [FAISS], which stands for Facebook AI Similarity Search. FAISS is utilized here due to its proficiency in managing high-dimensional vectors and facilitating rapid similarity searches, which are crucial for retrieving relevant information from the knowledge base.

Finally, the LLM (Llama-7B) uses this structured vector database to generate recommendations. The LLM analyzes the embeddings, correlates them with the user's health condition through a prompt, and produces pertinent recommendations, which could range from medical advice to specific treatment options.

1) Data Preprocessing for CNN model

Prior to the utilization of convolutional neural network (CNN) models for the classification of chest X-ray images, preprocessing is pivotal. This phase conditions the images into a format conducive for CNNs to effectively learn complex features and patterns, a prerequisite for precise classification. The preprocessing steps implemented are detailed as follows:

1) **Image Loading and Resizing:** Each selected image is loaded and resized to a uniform target dimension of 224x224 pixels, which matches the input size expected by the CNN models.
2) **Converting Images to NumPy Arrays:** Conversion of images to NumPy arrays is performed to facilitate efficient storage of pixel values, offering a form well-suited for subsequent CNN processing. The adjusted images are transformed into these arrays.
3) **Normalization:** The pixel values within the images are normalized to a range of [0, 1] by division with 255. Such normalization is integral to improving the generalization capabilities of the models and expediting the training phase.
4) **Expanding Dimensions:** To prepare the images for batch processing by CNNs, we expand the dimensions of the image arrays accordingly, catering to the batch input specifications of the deployed models.

2) Data Preprocessing for LLM Model

The preprocessing for our large language model (LLM) involves the creation of a vector database, integrating diverse data formats such as PDFs, CSV files, and scraped web content. This data is processed into textual embeddings, capturing the statistical probability and relational nuances of word occurrences. The semantic integrity of the sentences is paramount, as our chosen LLM model relies on it to generate contextually relevant responses. Employing FAISS—a library for efficient similarity search and clustering of high-dimensional vectors—we construct a vector database to act as the knowledge base (KB) for the LLM. The preprocessing encompasses the following key stages:

i. **Document Loading:** A document loader is configured to ingest all relevant PDFs pertaining to the medical context.
ii. **Text Splitting:** The loaded documents are segmented into smaller text chunks for more manageable processing, with each segment spanning approximately 500 characters and sharing an overlap of 50 characters with the subsequent segment.
iii. **Text Embeddings:** These text segments are then encoded into numerical embeddings utilizing HuggingFace's sentence-transformer models, specifically all-MiniLM-L6-v2.
iv. **Vector Database (FAISS) Creation:** A vector database is constructed from the embeddings, enabling rapid similarity searches within the corpus of text segments.
v. **Database Saving:** The resultant vector database is preserved, allowing for the embeddings and their corresponding text segments to be readily available for future recommendation and similarity search tasks without reprocessing.

E. Prompting Strategy

Within the scope of our system, we employ a few-shot prompting strategy, a method particularly suited for large language models (LLMs) like 'Llama-2-7B-chat-GGML'. Few-shot learning enables LLMs to grasp new tasks by considering only a small set of examples, or "shots". This approach leverages the natural language understanding capabilities of the model to generate text based on a limited number of illustrative prompts. To ensure the effectiveness and reliability of our few-shot prompting strategy, we adhere to the following prompt engineering policy:

- **Contextual Relevance:** Prompts are generated dynamically based on the patient's diagnosis, ensuring that the model's response is tailored to the specific health condition identified. This contextual relevance is crucial for generating accurate and useful information.

- **Comprehensive and Specific Requests:** Prompts are designed to request detailed information, including specific medications, dosages, and management advice for diagnosed conditions. This specificity helps guide the LLM to provide thorough and actionable recommendations.

- **General Health Tips:** In cases where no disease is detected, prompts are structured to request general health tips. This ensures that the model can still provide valuable information even in the absence of a specific diagnosis.

- **Knowledge Base (KB) Inference:** The LLM references a robust knowledge base to formulate responses to the prompts. The KB is regularly updated with medical literature, guidelines, and authoritative sources to ensure that the information provided is current and evidence-based.

- **Accuracy and Reliability:** The model is designed to prioritize accurate and useful responses. In scenarios where the LLM cannot ascertain a confident answer, it refrains from fabricating responses. This policy is

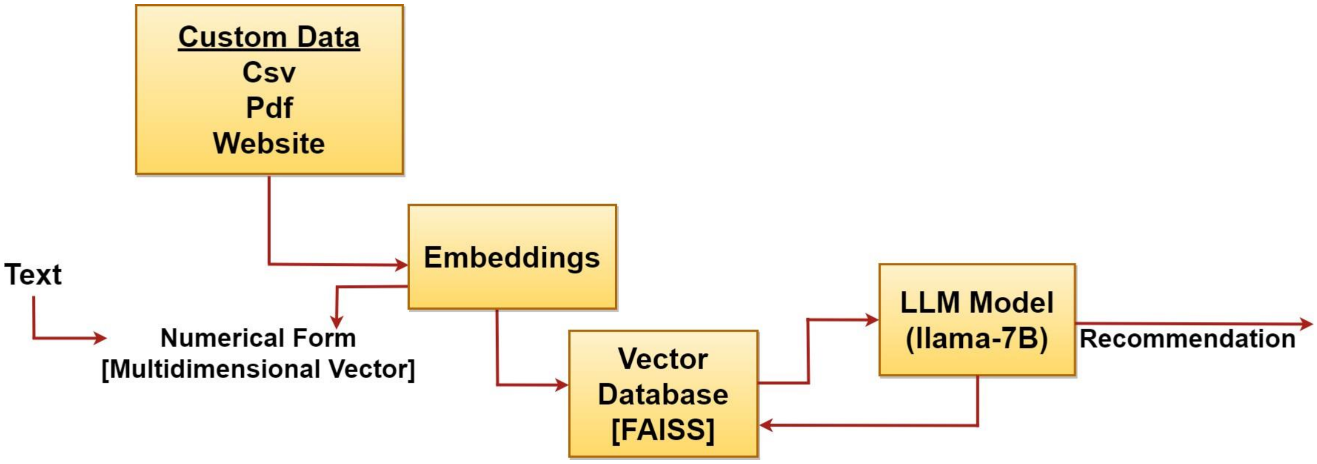

Figure 6. The detailed stages of data preprocessing for CNN and LLM models.

crucial for maintaining the reliability and trustworthiness of the information provided to patients.

Continuous Improvement: The prompt generation algorithm and the LLM's performance are continuously monitored and refined based on feedback and advancements in medical knowledge and AI capabilities. This iterative improvement ensures that the system remains effective and relevant over time.

The generation of prompts is crucial in guiding the model to provide the desired output. To structure these prompts effectively, we designed the following pseudocode, which is dynamically executed based on the diagnosis:

Algorithm 1 Prompt Generation Algorithm

```
if result is in [”Pneumonia”, ”COVID-19”] then
    prompt = ”As a patient recently diagnosed with ” result”,
I am looking for expert medical advice on how to manage this condition.
Could you provide detailed information on the specific medications recommended for ” result ”,
including their names, dosages, and any important instructions or precautions?
Furthermore, I would appreciate any additional professional advice or guidelines
on lifestyle adjustments, dietary considerations, and preventive measures to manage and improve my health condition.”
else
    prompt = ”I have been informed that I do not have any diseases.
However, I am interested in maintaining good health.
Could you please provide me with a valuable health tip or general advice to enhance my overall well-being?”
end if
```

Based on the algorithm (1), if a patient's chest X-ray analysis results in a diagnosis of Pneumonia or COVID-19, the generated prompt solicits comprehensive information regarding medication and management advice. Conversely, if no disease is detected, the prompt inquires about general health tips.

4. ExpeRIMenTal Analysis

Our experimental study aims to evaluate the performance of an integrated diagnostic system utilizing both convolutional neural networks (CNNs) and a large language model (LLM). The system's primary function is to analyze chest X-ray images to detect the presence of respiratory conditions, specifically Pneumonia and COVID-19, and provide corresponding health recommendations.

To begin, we utilized three CNN architectures renowned for their image classification efficacy: InceptionV3, MobileNetV2 trained for 200 epochs, and NASNet. The accuracy of these CNN models in classifying X-ray images into the aforementioned categories is critical to the success of the subsequent LLM recommendation process. The evaluation metrics used to assess the performance of the CNN models included:

- **Accuracy**: The proportion of correctly classified images out of the total number of images.
- **Precision**: The proportion of true positive results in the total predicted positive results.
- **Recall**: The proportion of true positive results out of the actual positive cases.
- **F1 Score:** The harmonic mean of precision and recall, provides a single metric that balances the two.

The accuracy of these models in classifying X-ray images into categories—normal, Pneumonia, or COVID-19—is critical to the success of the subsequent LLM recommendation process. The precision of the LLM’s health advice, based on the diagnosis provided by the CNNs, was also a vital aspect of the system’s overall evaluation. In our study, we utilized Google Colab notebooks to train our models, taking advantage of the robust and accessible computing resources they offer. By leveraging Colab’s freely accessible GPUs, specifically the Tesla T4 GPUs, we could expedite the training of our complex convolutional neural network (CNN) models.

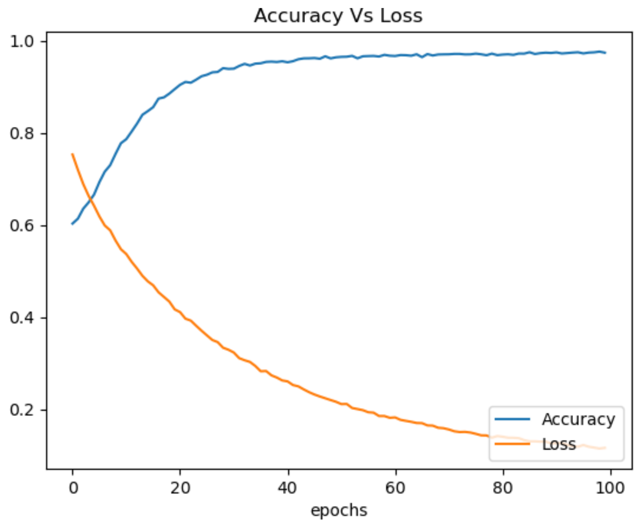

Figure 7. Accuracy versus loss graph for the InceptionV3 model during validation.

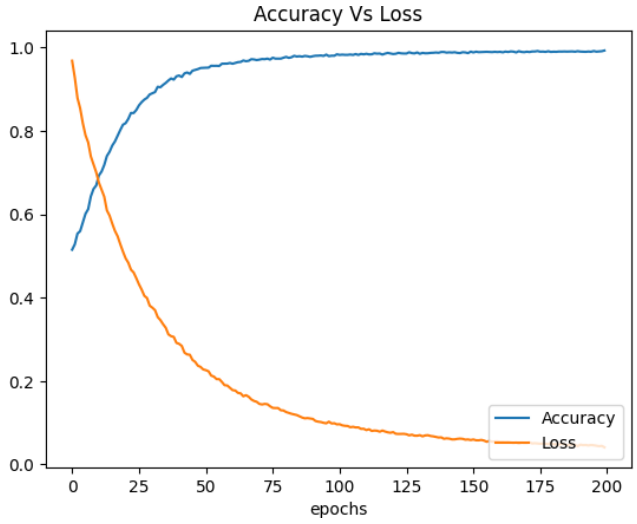

Figure 8. MobileNetV2 (200 epochs) model accuracy versus loss graph.

Figures 7, 8, and 9 show the accuracy versus loss curves for the InceptionV3, MobileNetV2, and NASNet models, respectively. Notably, the InceptionV3 model achieved an impressive accuracy of 92.85% and a loss of 17.09%. The MobileNetV2 model reached an accuracy of 91.88% after 200 epochs, with a loss of 20.39%. These results are indicative of the robust diagnostic capabilities the CNN models offer to the overall system. The plots reveal a common trend: as the number of epochs increases, the accuracy stabilizes, and the loss decreases, which is characteristic of a well-fitting model. The analysis of these curves provides insights into the models’ learning dynamics over time and underscores the importance of adequate training in achieving high performance. The results obtained from the CNN models form the basis for the subsequent phase where the LLM—’Llama-2-7B-chat-GGML’—takes over. Based on the categorized images, the LLM generates tailored recommendations, drawing from its training and the knowledge base built from diverse medical data. The experimental analysis thus serves a dual purpose: validating the CNN models’ classification accuracy and setting the stage for the LLM’s effective recommendation generation.

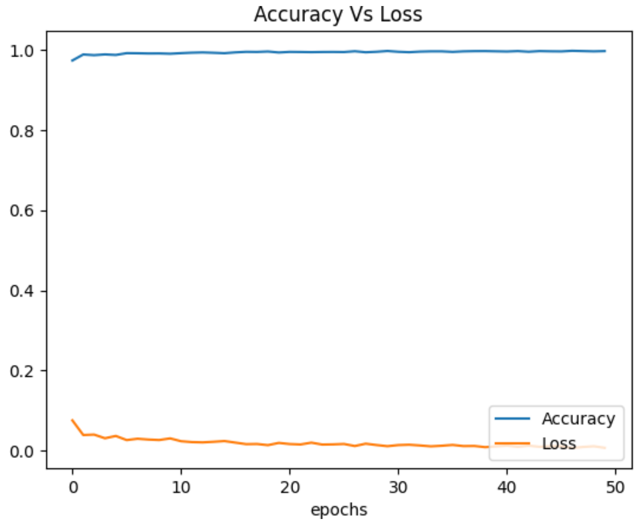

Figure 9. Accuracy versus loss graph for the NASNet model during validation.

In our study, we employed three different CNN architectures to classify chest X-ray images into normal, pneumonia, and COVID-19 categories. The performance of each model was rigorously evaluated based on several metrics, including accuracy, loss, precision, recall, and the F1 score. Table III provides a concise summary of these results.

Analyzing the results (Table III), NasNet outperforms the other models with the highest accuracy of 95.92% and an impressive precision of 96.64%. It also demonstrates the lowest loss rate, indicating its robustness in X-ray image classification tasks. While InceptionV3 shows a comparable performance in terms of precision and recall, its slightly higher loss rate suggests a minor susceptibility to overfitting compared to NasNet. Conversely, MobileNetV2, despite a modest dip in accuracy and precision, presents a balanced profile that is potentially less prone to overfitting, evidenced

TABLE III. Summary of CNN Models Result Analysis and Performances

CNN Model	Accuracy	Loss	Precision	Recall	F1 score
InceptionV3	92.85%	17.09%	94.03%	92.20%	93.11%
MobileNetV2	91.88%	20.39%	91.44%	90.25%	90.84%
NasNet	95.92%	14.22%	96.64%	95.92%	96.28%

by its consistent performance across metrics.

In evaluating the performance of our CNN models, we place significant emphasis on understanding their classification accuracy and potential misclassification trends. Confusion matrices for each of the models—InceptionV3, MobileNetV2 trained for 200 epochs, and NASNet—provide valuable insights into the true positive, true negative, false positive, and false negative rates, which are instrumental in gauging the clinical applicability of these models.

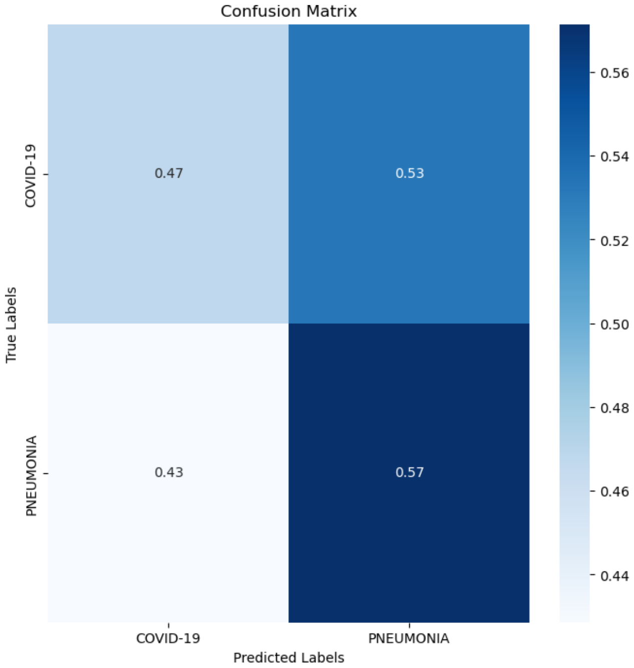

Figure 10. Confusion matrix for the InceptionV3 model.

Figure 10 details the confusion matrix for the InceptionV3 model, showing a balance of 47% true positives to 53% false negatives, and 43% false positives to 57% true negatives. This suggests a higher sensitivity in detecting pneumonia.

Similarly, Figure 11 exhibits the confusion matrix for MobileNetV2, where the distinction between true positives and false negatives is more pronounced at 47% and 37%, respectively, indicating a propensity towards more conservative classification.

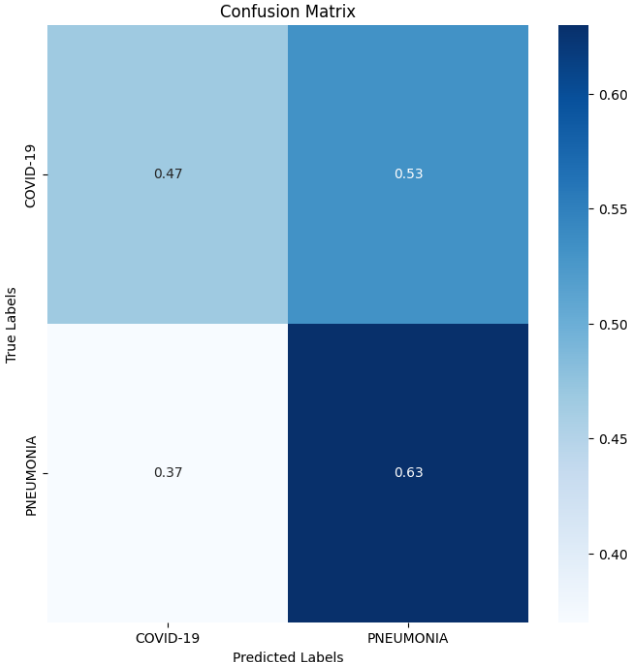

Figure 11. Confusion matrix for the MobileNetV2 model after 200 epochs of training.

Lastly, the NASNet model's confusion matrix, as seen in Figure 12, reveals 35% true positives against a considerable 65% false positives, suggesting a potential inclination towards overpredicting the presence of pneumonia. Through these matrices, we can deduce the strengths and weaknesses of each CNN model in diagnosing respiratory diseases from X-ray images. The comprehensive analysis of true positive and false negative rates is critical in further refining these models to reduce diagnostic errors in real-world medical settings.

The efficacy of our prompt generation algorithm is a significant contributor to the overall functionality of the 'Llama-2-7B' model, which is tasked with classifying diseases from chest X-ray images. We assessed the performance based on two critical criteria: the relevance of the prompts generated by the algorithm and the quality of the responses produced by the model.

The data presented in Tables IV and V illustrate the prompt relevance and response efficacy of the LLM model

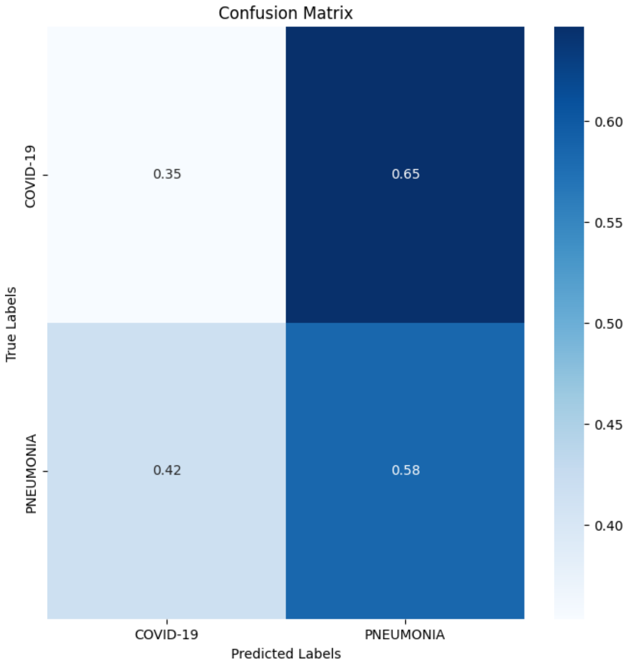

Figure 12. Confusion matrix for the NASNet model.

TABLE IV. Prompt, Response, and First Inference Time for Pneumonia

Prompt	As a patient who has received a diagnosis of pneumonia, I am seeking expert guidance on the recommended medications and advice for my condition. Please provide detailed information on the specific medications I should take, including their names, dosages, and any important instructions. Additionally, I would greatly appreciate any professional advice or recommendations on what actions I should take to effectively manage my pneumonia.
Response	[Response abbreviated for brevity; the actual response provided by the LLM lists several supportive measures for managing pneumonia.]
Inference Time	165.6 seconds

in two diagnostic scenarios. Notably, there is a reduction in inference time from the first to the second scenario, indicating an increase in processing efficiency. This improvement suggests that our model is becoming more adept at generating relevant responses in a shorter time frame, a crucial factor in providing timely medical advice.

5. Discussion

A. Optimization of CNN Architectures

Our experimental results confirm that with appropriate tuning, CNNs such as InceptionV3 and MobileNetV2 exhibit high levels of accuracy in image classification tasks. Notably, NasNet demonstrated superior performance with an accuracy of 95.92%, a precision rate of 96.64%, and the lowest loss rate of 14.22%. These results underscore NasNet's nuanced feature extraction capabilities and robustness in distinguishing between normal, pneumonia, and COVID-19 cases. However, the challenge of overfitting remains prevalent, as indicated by the slightly higher loss rates in InceptionV3 and MobileNetV2. This necessitates ongoing refinement of regularization techniques and data augmentation strategies to enhance model generalization.

TABLE V. Prompt, Response, and Second Inference Time for COVID-19

Prompt	As a patient who has received a diagnosis of COVID-19, I am seeking expert guidance on the recommended medications and advice for my condition. Please provide detailed information on the specific medications I should take, including their names, dosages, and any important instructions. Additionally, I would greatly appreciate any professional advice or recommendations on what actions I should take to effectively manage my COVID-19.
Response	[Response abbreviated for brevity; the actual response from the LLM includes a disclaimer about its limitations as a language model and a reminder to follow official health guidelines.]
Inference Time	111.9 seconds

B. Employment of LLMs for Health Recommendations

The study explored the effectiveness of LLMs in generating health recommendations. The 'Llama-2-7B' model was instrumental in providing contextually relevant advice with inference times reflecting efficient deployment. For instance, the inference time for generating detailed recommendations for pneumonia was 165.6 seconds, while for COVID-19 it was reduced to 111.9 seconds, demonstrating improved processing efficiency. The quality of the responses, measured against professional health guidelines, suggests that while LLMs are capable of generating general advice, their utility is significantly enhanced when coupled with expert oversight. This finding corroborates related studies advocating for a hybrid approach that combines AI with human expertise to ensure the accuracy and reliability of health recommendations.

C. Integration of AI in Clinical Workflow

Our findings reveal that AI models can significantly expedite the diagnostic process, as evidenced by the high accuracy and precision rates obtained in our experiments. For example, NasNet achieved an F1 score of 96.28%, indicating its potential as a reliable adjunct to human diagnosticians. However, the nuanced nature of medical practice, which extends beyond the rigid classifications of AI models, underscores the necessity for these tools to operate within a framework that respects the complexity of patient care. The integration of AI must be approached with caution, ensuring that AI outputs are used to support, rather than replace, clinical judgment.

D. Limitations and Future Work

Our study is not without limitations. The size and diversity of the dataset, potential biases in model training, and the interpretability of model decisions are critical areas that require further investigation. For instance, while our dataset comprised 8,118 chest X-ray images, expanding this dataset to include a broader range of patient demographics and conditions could enhance model robustness and generalizability. Additionally, the potential biases inherent in training data need to be addressed to prevent skewed predictions.

The responses generated by the LLM also necessitate validation from medical professionals to ensure accuracy and reliability. Future research will aim to address these limitations by expanding the dataset, exploring techniques to mitigate biases, and developing methods to enhance the explainability of AI models. Furthermore, the integration of multimodal data sources, such as electronic health records, could enrich the LLM’s knowledge base, leading to more nuanced and comprehensive recommendations. Developing transparent models that provide clear rationale for their predictions will be crucial in gaining clinician trust and facilitating seamless integration into healthcare settings.

6. ConclusIOn

Our study demonstrates the promising potential of integrating convolutional neural networks (CNNs) with large language models (LLMs) to enhance the accuracy and comprehensiveness of AI-assisted medical diagnostics and recommendations. The experimental analysis revealed that advanced CNN architectures such as InceptionV3, MobileNetV2, and NasNet can effectively classify chest X-ray images to detect respiratory conditions, specifically Pneumonia and COVID-19, with high precision and accuracy. Among these, NasNet showed superior performance, highlighting its capability for nuanced feature extraction and robust classification.

The integration of the LLM, specifically the ’Llama-2-7B’ model, for generating health recommendations further extends the utility of the diagnostic system. The LLM was able to provide contextually relevant and detailed medical advice, demonstrating efficient inference times and high-quality responses. This dual-component system, combining the strengths of CNNs in image classification and LLMs in natural language processing, offers a comprehensive approach to managing respiratory diseases. However, the study also identified several areas requiring further research and development. The potential for overfitting in CNN models, the need for a more diverse and extensive dataset, and the necessity for validation of LLM-generated responses by medical professionals are critical considerations for future work. Additionally, addressing biases in model training and enhancing the interpretability of AI decisions are paramount to ensure reliable and trustworthy AI applications in healthcare.

Future research will focus on expanding the dataset to include a wider range of patient demographics and conditions, exploring advanced techniques to mitigate biases, and developing transparent AI models that provide clear rationales for their predictions. Integrating multimodal data sources, such as electronic health records, will also be pursued to enrich the LLM’s knowledge base and enable more nuanced recommendations.

In conclusion, the integration of CNNs for accurate disease detection with LLMs for tailored health recommendations represents a significant advancement in AI-assisted medical diagnostics. This approach not only improves diagnostic accuracy but also enhances the quality of patient care through personalized and actionable medical advice. Realizing the full potential of this integrated system will require continuous refinement and a collaborative effort between AI researchers and healthcare professionals.

RefeRences

[1] K. K. Bressem, L. C. Adams, R. A. Gaudin, D. Tro¨ltzsch, B. Hamm, M. R. Makowski, C.-Y. Schu¨le, J. L. Vahldiek, and S. M. Niehues, “Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports,” *Bioinformatics*, vol. 36, no. 21, pp. 5255–5261, 2020.

[2] M. Sallam, “The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations,” *medRxiv*, pp. 2023–02, 2023.

[3] M.-J. Sanaei, M. S. Ravari, and H. Abolghasemi, “Chatgpt in medicine: Opportunity and challenges,” *Iranian Journal of Blood and Cancer*, vol. 15, no. 3, pp. 60–67, 2023.

[4] M. Salehi, R. Mohammadi, H. Ghaffari, N. Sadighi, and R. Reiazi, “Automated detection of pneumonia cases using deep transfer learning with paediatric chest x-ray images,” *The British journal of radiology*, vol. 94, no. 1121, p. 20201263, 2021.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[6] E. Khan, M. Z. U. Rehman, F. Ahmed, F. A. Alfouzan, N. M. Alzahrani, and J. Ahmad, “Chest x-ray classification for the detection of covid-19 using deep learning techniques,” *Sensors*, vol. 22, no. 3, p. 1211, 2022.

[7] C. Wang, C. Xu, Y. Zhang, and P. Lu, “Diagnosis of chest pneumonia with x-ray images based on graph reasoning,” *Diagnostics*, vol. 13, no. 12, p. 2125, 2023.

[8] A. Mabrouk, R. P. D´ıaz Redondo, A. Dahou, M. Abd Elaziz, and M. Kayed, “Pneumonia detection on chest x-ray images using

ensemble of deep convolutional neural networks," *Applied Sciences*, vol. 12, no. 13, p. 6448, 2022.

[9] T. Le Dinh, S.-H. Lee, S.-G. Kwon, and K.-R. Kwon, "Covid-19 chest x-ray classification and severity assessment using convolutional and transformer neural networks," *Applied Sciences*, vol. 12, no. 10, p. 4861, 2022.

[10] X. Huang, J. Lian, Y. Lei, J. Yao, D. Lian, and X. Xie, "Recommender ai agent: Integrating large language models for interactive recommendations," *arXiv preprint arXiv:2308.16505*, 2023.

[11] H. Ding, Y. Ma, A. Deoras, Y. Wang, and H. Wang, "Zero-shot recommender systems," *arXiv preprint arXiv:2105.08318*, 2021.

[12] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, and W. X. Zhao, "Large language models are zero-shot rankers for recommender systems," *arXiv preprint arXiv:2305.08845*, 2023.

[13] Z. Cui, J. Ma, C. Zhou, J. Zhou, and H. Yang, "M6-rec: Generative pretrained language models are open-ended recommender systems," *arXiv preprint arXiv:2205.08084*, 2022.

[14] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, "Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5)," in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 299–315.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] M. Karabacak and K. Margetis, "Embracing large language models for medical applications: Opportunities and challenges," *Cureus*, vol. 15, no. 5, 2023.

[18] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *arXiv preprint arXiv:2212.13138*, 2022.

[19] H. Fei, Y. Ren, Y. Zhang, D. Ji, and X. Liang, "Enriching contextualized language model from knowledge graph for biomedical information extraction," *Briefings in bioinformatics*, vol. 22, no. 3, p. bbaa110, 2021.

[20] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, "Chatcad: Interactive computer-aided diagnosis on medical image using large language models," *arXiv preprint arXiv:2302.07257*, 2023.

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozie`re, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[22] Y. Hou, S. Mu, W. X. Zhao, Y. Li, B. Ding, and J.-R. Wen, "Towards universal sequence representation learning for recommender systems," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 585–593.

[23] Y. Hou, Z. He, J. McAuley, and W. X. Zhao, "Learning vector-quantized item representation for transferable sequential recommenders," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 1162–1171.

[24] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang, "Chat-rec: Towards interactive and explainable llms-augmented recommender system," *arXiv preprint arXiv:2303.14524*, 2023.

[25] J. Liu, C. Liu, R. Lv, K. Zhou, and Y. Zhang, "Is chatgpt a good recommender? a preliminary study," *arXiv preprint arXiv:2304.10149*, 2023.

[26] J. Zhang, R. Xie, Y. Hou, W. X. Zhao, L. Lin, and J.-R. Wen, "Recommendation as instruction following: A large language model empowered recommendation approach," *arXiv preprint arXiv:2305.07001*, 2023.

[27] D. Sileo, W. Vossen, and R. Raymaekers, "Zero-shot recommendation as language modeling," in *European Conference on Information Retrieval*. Springer, 2022, pp. 223–230.

[28] Y. Shu, H. Gu, P. Zhang, H. Zhang, T. Lu, D. Li, and N. Gu, "Rah! recsys-assistant-human: A human-central recommendation framework with large language models," *arXiv preprint arXiv:2308.09904*, 2023.

[29] C. Wang, D. Chen, L. Hao, X. Liu, Y. Zeng, J. Chen, and G. Zhang, "Pulmonary image classification based on inception-v3 transfer learning model," *IEEE Access*, vol. 7, pp. 146 533–146 541, 2019.

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[31] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," 2018.

[32] A. Asraf and Z. Islam, "Covid19, pneumonia and normal chest x-ray pa dataset. mendeley data v1 (2021)," 2021.