AI System for Disease Prediction and Pneumonia detection

Submitted for

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING CSET301

Submitted by:

(E23CSEU1166) Robin Singh

(E23CSEU1110) Advitiya Arya

(E23CSEU1156) Sourabh Pal

Submitted to

Dr. Shwetang Dubey

Jan-May 2024

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

INDEX

1. Abstract

This project presents an AI-based healthcare system designed to predict diseases from user-described symptoms and detect pneumonia using chest X-ray images. By integrating both machine learning and deep learning techniques, the system offers a hybrid diagnostic solution for early and accessible health screening.

Users can input symptoms along with lifestyle factors such as smoking or medical history, and the system predicts the top three most likely diseases while also providing precautions and recommendations. For users who suspect respiratory issues, the system allows uploading of chest X-ray images, which are analyzed using a deep learning model to detect signs of pneumonia. Both predictions are combined to present a final diagnosis summary through an intuitive web interface.

This solution aims to improve early diagnosis and raise health awareness, particularly in areas with limited access to healthcare professionals or diagnostic infrastructure.

2. Introduction

Accurate and timely diagnosis is a cornerstone of effective healthcare. However, in many regions, access to qualified medical personnel and diagnostic tools is limited. To bridge this gap, artificial intelligence offers promising solutions that can assist in early disease detection and self-assessment.

This project introduces a hybrid AI system that combines text-based symptom analysis and image-based disease detection. The system is designed to serve as an accessible tool for users to gain insights into their health conditions by simply describing their symptoms and optionally uploading chest X-rays. A machine learning model processes the symptom data to predict likely diseases, while a deep learning model analyzes chest X-rays to detect pneumonia—a common but potentially serious respiratory condition.

The system is deployed through a web-based interface, providing users with personalized predictions, precautionary advice, and guidance, all without requiring immediate access to a hospital or clinic. This integration of machine learning and deep learning enhances diagnostic reliability and expands accessibility for preventive healthcare.

3. Related Work

AI-driven medical diagnostics have been extensively researched in recent years, leading to the development of various systems for disease prediction and image-based analysis:

Symptom-based prediction systems use machine learning models trained on datasets containing symptoms and disease mappings. Previous works have demonstrated success in predicting illnesses by analyzing the correlation between common symptoms and medical conditions.

Lifestyle factor integration has been shown to improve prediction accuracy. Studies have explored combining patient-reported symptoms with risk factors like age, smoking habits, and existing conditions (e.g., diabetes, hypertension) to create more personalized diagnostic outputs.

Medical image analysis using deep learning has particularly advanced pneumonia detection. Several projects have utilized convolutional neural networks (CNNs) trained on large public datasets of chest X-rays to achieve high accuracy in binary classification (e.g., pneumonia vs. normal).

Hybrid diagnostic systems that incorporate both symptom data and imaging analysis have recently emerged, allowing for cross-validation of diagnoses and more robust health assessments. These systems help mitigate false positives and improve the confidence of predictions.

Building on these foundations, this project provides an integrated, user-friendly platform that combines these technologies into a unified diagnostic tool accessible through a web interface.

4. Problem Statement

In the current healthcare landscape, early and accurate disease diagnosis remains a challenge, especially in regions with limited access to medical professionals and diagnostic tools. Patients often delay medical consultations due to uncertainty regarding the severity of their symptoms or lack of awareness, resulting in worsened health outcomes. Additionally, existing AI-based diagnostic systems are often limited to either symptom analysis or image-based diagnosis but rarely integrate both modalities for a comprehensive evaluation.

There is a critical need for a user-friendly, AI-driven platform that can assess a user's health risks by analyzing both textual symptom descriptions and medical imaging, such as chest X-rays. Moreover, incorporating lifestyle factors such as smoking habits, diabetes, and hypertension can significantly enhance diagnostic accuracy and personalize recommendations. Such a system should also provide preventive measures and medical advice to empower users toward better health decisions.

5. Contribution

This project presents a hybrid AI-powered health diagnostic system that combines machine learning and deep learning for disease prediction and personalized recommendations.

Symptom & Lifestyle-Based Prediction:
A Random Forest classifier, trained on user-inputted symptoms and lifestyle factors, predicts the top three probable diseases along with confidence scores, medical descriptions, and preventive tips.

Medical Image Analysis:
A Convolutional Neural Network (CNN) processes chest X-rays to detect high-risk diseases like Pneumonia and Tuberculosis, supporting binary classification (disease vs. no disease).

Integrated Diagnosis Summary:
The system merges results from both models, considering symptom analysis, lifestyle risks, and optional X-ray confirmation to provide a comprehensive diagnosis summary.

6. Work-Load Division

1. Robin

Responsibilities:

- **Model Training: CNN**
  - Train and fine-tune the model using chest X-ray datasets.
  - Perform image preprocessing (resize, normalize, augment).
  - Evaluate CNN model using accuracy/loss curves and sample outputs.
- **Report Sections:**
  - Medical Image Analysis with CNN
  - Dataset preprocessing and augmentation
  - Visual examples and discussion on disease detection via X-ray

2.Advitiya

Responsibilities:

- **Model Training: Random Forest**
  - Model Training: Random Forest
  - Train the Random Forest classifier using symptom and lifestyle data.
  - Preprocess textual symptom data and encode lifestyle attributes.
  - Optimize and evaluate model performance.
- **Report Sections:**
  - Symptom & Lifestyle-Based Prediction
  - Text preprocessing and feature engineering methods
  - Random Forest training and evaluation
  - Result interpretation and visualization
  - Contribution to diagnosis summary

3. Sourabh

Responsibilities:

- **Model Integration & Testing**
  - Support testing and validation of both Random Forest and CNN models.
  - Write integration logic that combines symptom-based and image-based predictions.
  - Handle confidence score merging and decision-making logic.
- **Report Sections:**
  - Final diagnosis summary (merging predictions)
  - Model comparison and integration explanation
  - Web app backend structure (Flask)
  - Chart.js implementation and result visualization
  - Final deployment and testing report

7. Methodology

The system employs a hybrid AI approach combining machine learning and deep learning for health risk assessment.

4.1 Symptom & Lifestyle-Based Prediction

- **Text Preprocessing**: User symptoms are processed using TF-IDF vectorization to convert text into numerical form.
- **Feature Integration**: Lifestyle-related inputs (e.g., smoker, diabetes, hypertension) are combined with symptom vectors.
- **Classification**: A Random Forest classifier predicts the most probable diseases, returning the top 3 predictions with confidence scores, medical descriptions, general precautions, and custom recommendations based on user profile.

4.2 Medical Image Analysis (CNN)

- For high-risk diseases (e.g., Tuberculosis, Pneumonia), users can upload chest X-ray images.
- A Convolutional Neural Network (CNN) model analyzes the images to assist or verify disease prediction.
- The CNN uses two convolutional and pooling layers followed by fully connected layers for binary classification (e.g., disease vs. no disease).

4.3 Final Diagnosis Summary

The system merges insights from:

- Symptom-based prediction
- Lifestyle risk scoring
- Optional image-based confirmation (for high-risk conditions) to generate a comprehensive diagnosis summary.

4b.) Dataset Development

4.1 Symptom & Lifestyle Dataset

- Collected from open-source medical repositories and cleaned to include:
  - Disease name
  - Associated symptoms
  - Lifestyle factors (binary)
  - Disease stage (numeric)
- Used to train the Random Forest classifier with TF-IDF features and encoded lifestyle attributes.

4.2 Medical Imaging Dataset

- Chest X-ray images sourced from publicly available datasets (e.g., NIH Chest X-ray or COVIDx) for training the CNN model.
- Images are resized to 150×150 pixels, normalized, and labeled for binary classification tasks.

4c) Web Application

The frontend is built using **HTML**, **CSS**, and **JavaScript**, served via a **Flask backend**.

- **Form Page**: Collects user inputs like symptoms, age, gender, lifestyle habits, and optionally a scan image.
- **Results Page**: Displays predictions with confidence, descriptions, precautions, and custom health tips.
- **Visualization**: Uses chart.js to graphically represent prediction probabilities.

8. Experimental Result

```
Training Random Forest model...
Accuracy: 1.0
Classification Report:
```

	precision	recall	f1-score	support
(vertigo) Paroymsal Positional Vertigo	1.00	1.00	1.00	18
AIDS	1.00	1.00	1.00	30
Acne	1.00	1.00	1.00	24
Alcoholic hepatitis	1.00	1.00	1.00	25
Allergy	1.00	1.00	1.00	24
Arthritis	1.00	1.00	1.00	23
Bronchial Asthma	1.00	1.00	1.00	33
Cervical spondylosis	1.00	1.00	1.00	23
Chicken pox	1.00	1.00	1.00	21
Chronic cholestasis	1.00	1.00	1.00	15
Common Cold	1.00	1.00	1.00	23
Dengue	1.00	1.00	1.00	26
Diabetes	1.00	1.00	1.00	21
Dimorphic hemmorhoids(piles)	1.00	1.00	1.00	29
Drug Reaction	1.00	1.00	1.00	24
Fungal infection	1.00	1.00	1.00	19
GERD	1.00	1.00	1.00	28
Gastroenteritis	1.00	1.00	1.00	25
Heart attack	1.00	1.00	1.00	23

Medical Conditions:

None reported

Disease Prediction Summary

Disease	Confidence	Risk Level
Fungal infection	21.0%	Low
Drug Reaction	10.0%	Low
Heart attack	8.0%	Low

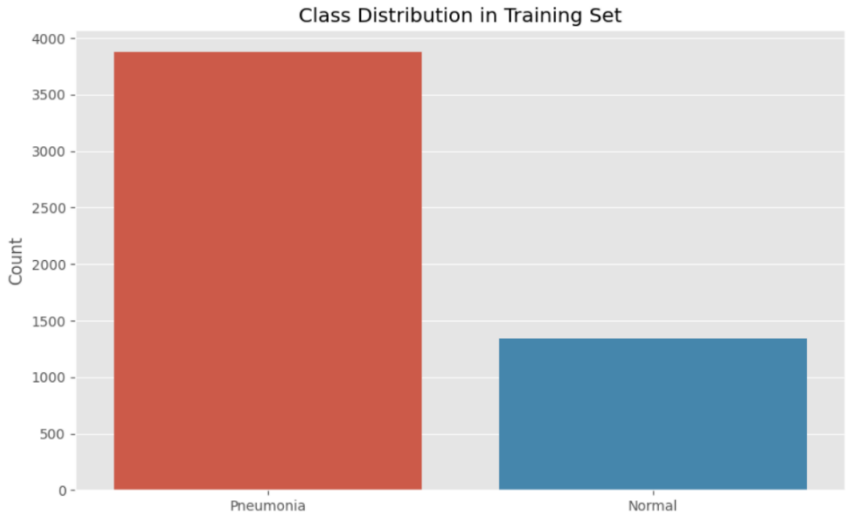
Disease Details: Fungal infection

Description:

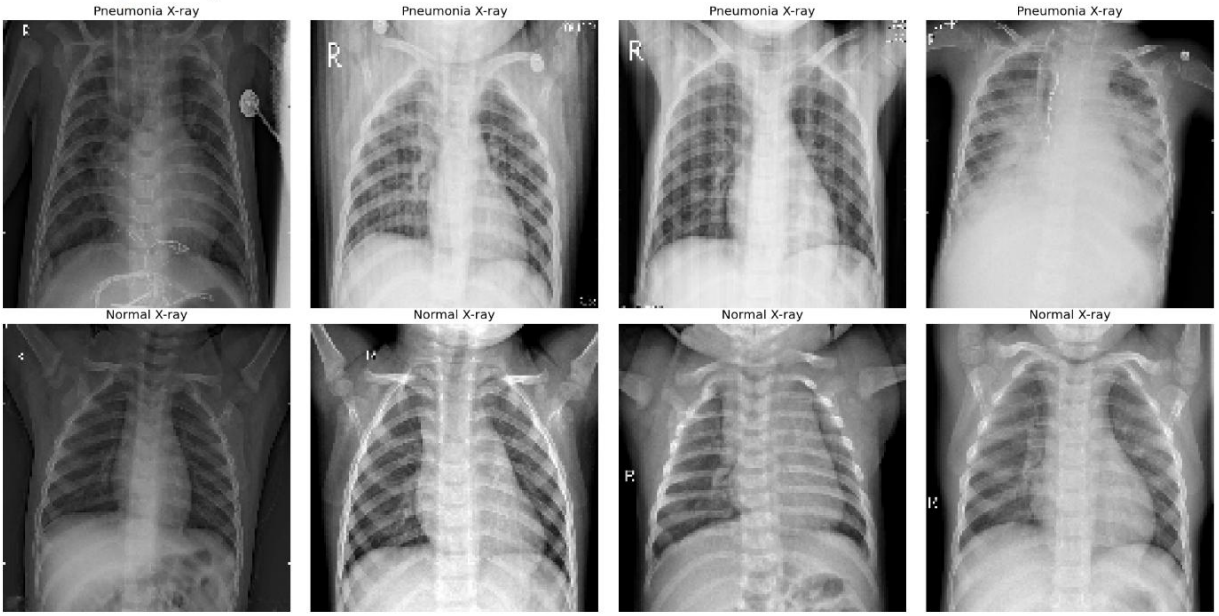
In humans, fungal infections occur when an invading fungus takes over an area of the body and is too much for the immune system to handle. Fungi can live in the air, soil, water, and plants. There are also some fungi that live naturally in the human body. Like many microbes, there are helpful fungi and harmful fungi.

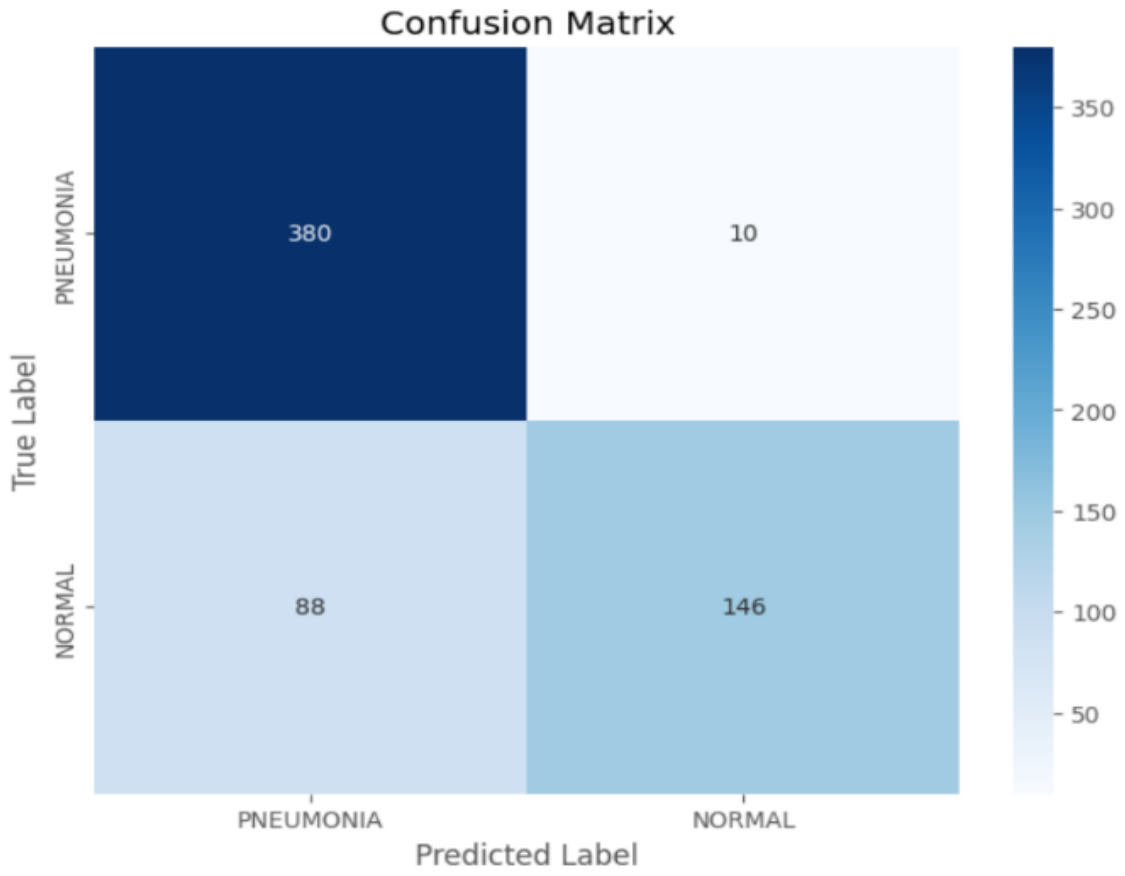
Precautions:

- bath twice
- use detol or neem in bathing water
- keep infected area dry
- use clean cloths

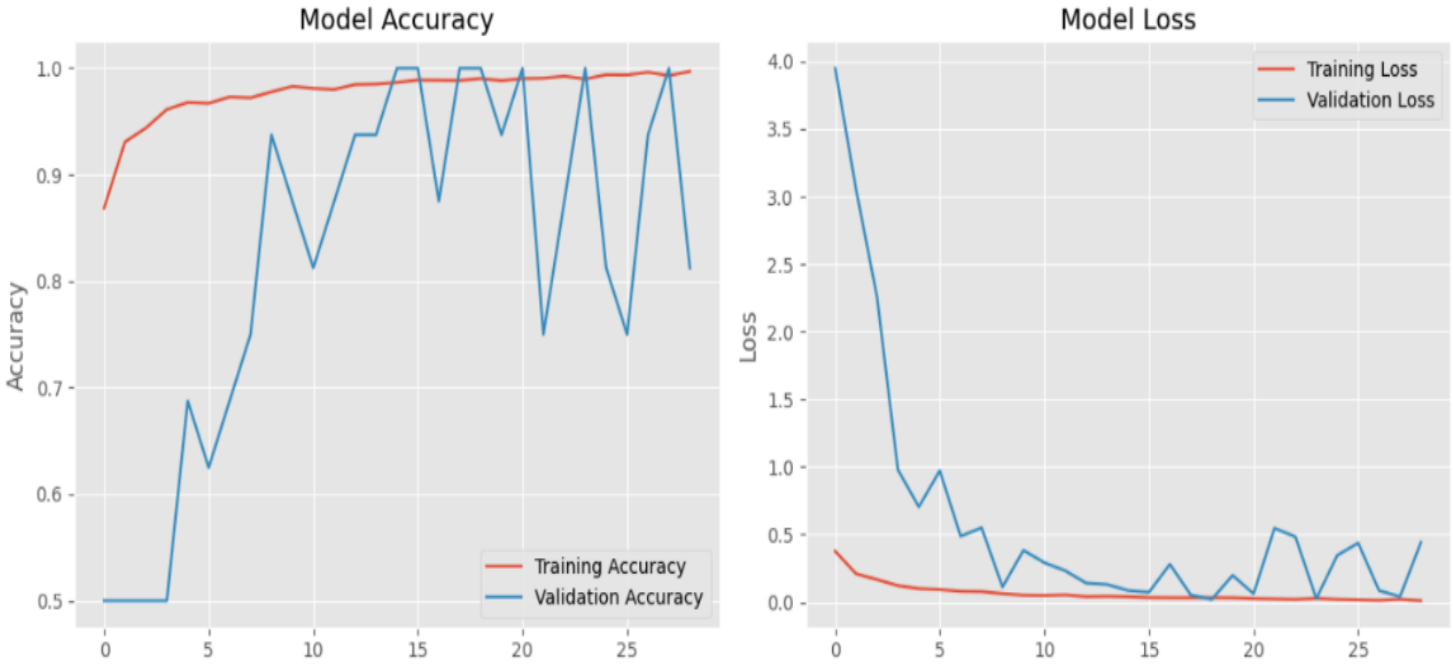



```
=== Visualizing Sample Images ===
```






```
=== Plotting Training History ===
```



9. Conclusion

This project demonstrates a hybrid AI-based Health Risk Assessment System that combines symptom analysis, lifestyle evaluation, and medical image interpretation for more accurate disease prediction. By integrating multiple input modalities, the system aims to support early detection and informed decision-making.

- **Data Limitations**: The symptom-lifestyle dataset may not capture rare or emerging diseases. Image datasets may lack diversity in demographics and image quality.
- **Binary Image Classification**: The CNN currently supports only binary outputs (e.g., TB vs. No TB), limiting its use for multi-class diagnoses.
- **Generalization**: The model might not perform well on unseen symptom patterns or complex comorbidities due to limited training data.
- **User Input Dependency**: Accuracy depends on the user entering correct and detailed symptoms.
- **Image Quality Variance**: Image analysis can be affected by upload quality, lighting, or resolution inconsistencies.

9. Refrences

- NIH Chest X-ray Dataset – https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

- Scikit-learn: Machine Learning in Python – Pedregosa et al., Journal of Machine Learning Research, 2011

- TensorFlow and Keras – Deep Learning Frameworks for CNN Implementation

- Open Source Medical Datasets – Collected from sources like SymCat, Disease Ontology, and WHO data repositories

10. Future Scope

- **Multi-class Image Classification**: Expand CNN models to detect multiple diseases like Lung Cancer, and COVID-19.
- **NLP-based Symptom Parsing**: Enhance symptom understanding using advanced language models for better user input interpretation.
- **Real-time Feedback Loop**: Integrate user feedback to continuously improve model predictions.
- **Wearable & Sensor Integration**: Include data from fitness trackers or medical sensors for real-time health monitoring.
- **Mobile Application Deployment**: Build a cross-platform mobile app for broader accessibility and convenience.
- **Doctor Dashboard**: Develop an admin panel for doctors to review cases, track trends, and update recommendations.