MACHINE LEARNING CS-324

Open Ended Lab

Prepared by:
Abdul Basit Siddiqui (CS-21105)
Saad Mustafa (CS-21113)
Muhammad Asad Shahab(CS-21126)

Presented to:
Miss Mahnoor Malik

Introduction

In medical diagnostics, machine learning techniques have opened new avenues for enhancing diagnostic accuracy and efficiency. This report presents a study focused on leveraging Support Vector Machines (SVM) and K Nearest Neighbors (KNN) algorithms for the automated classification of chest X-ray images to detect pneumonia.

Pneumonia remains a significant public health concern worldwide, necessitating timely and accurate diagnosis to guide appropriate medical interventions. This project demonstrates the efficacy of SVM and KNN in analyzing a dataset comprising X-ray images to distinguish between pneumonia-positive cases and normal conditions.

Data Collection

The first step was data collection. We obtained the dataset from Kaggle, which contained separate folders for Pneumonia and Normal images. Using Google Colaboratory, we imported the Kaggle dataset with the API command.

Data Preprocessing

Conversion to Pandas Dataframe

The set of images was converted to a Pandas data frame and saved in a CSV file.

RGB to Grayscale Conversion

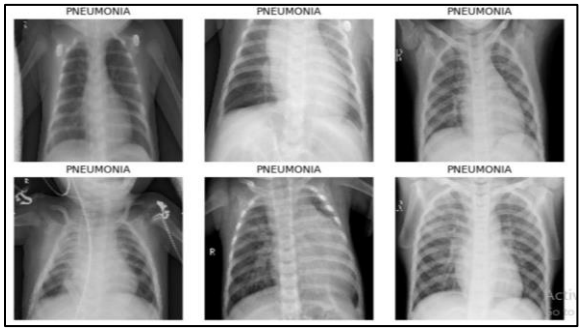
Each image was converted to grayscale, as color does not impact the prediction of pneumonia from X-ray images.

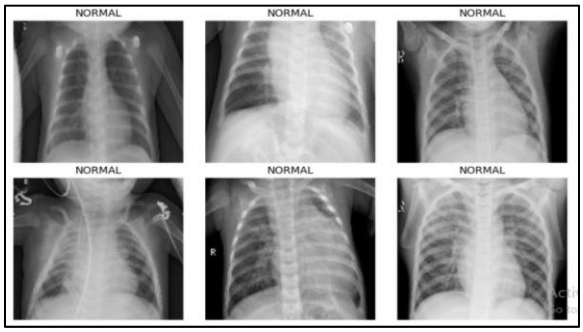
Image Resizing

Each image was resized to 150 x 150 pixels to make it feasible to train the model on such a large dataset.

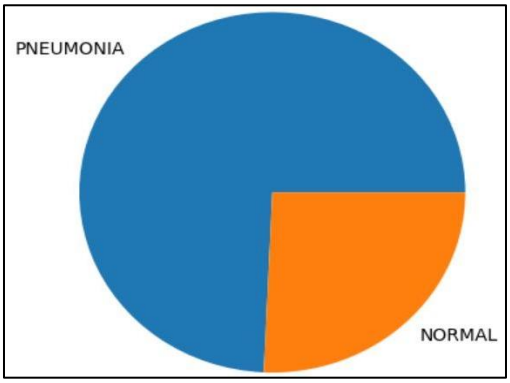
Exploratory Data Analysis

During our basic EDA, we resized high-resolution images to 150 x 150 pixels. Fortunately, the resized images retained sufficient clarity and detail, making them suitable for machine learning models.





Moreover, we checked the distribution of classes using a pie chart:



The above chart depicts that our dataset is highly imbalanced. Therefore, we balanced our training dataset using the Random Under Sampling technique. In this way, we were left with 1341 records of each class in our training dataset.

We also checked for some null values and fortunately, there were none.

Model Training

Given the size and complexity of our dataset, we opted for two machine learning algorithms: Support Vector Machines (SVM) and K Nearest Neighbors (KNN). We implemented these algorithms using Scikit-Learn and also developed them from scratch.

Model Description

K Nearest Neighbors (KNN) Classifier:

The K Nearest Neighbors (KNN) classifier is a simple, non-parametric algorithm used for classification. It classifies a data point based on the classes of its k-nearest neighbors in the feature space. For a new data point, the algorithm calculates the distance to all other points, identifies the k closest points, and assigns the class most common among them. This method is effective for pattern recognition and image analysis, as it makes no assumptions about the data distribution and relies on the similarity between data points for classification.

Support Vector Machine (SVM) Classifier:

The Support Vector Machine (SVM) classifier is a powerful, supervised learning algorithm used for classification tasks. It works by finding the optimal hyperplane that maximally separates data points of different classes in the feature space. SVM aims to identify the hyperplane that has the largest margin, meaning the greatest distance between the hyperplane and the nearest data points from each class, known as support vectors. By transforming the data into a higher-dimensional space using kernel functions, SVM can handle both linear and non-linear classification tasks effectively. It is widely used in various fields, including image recognition, bioinformatics, and text classification, due to its robustness and accuracy.

Results

Support Vector Machine (SVM) Classifier:

SKLEARN

	precision	recall	f1-score	support
-1	0.73	0.88	0.79	194
1	0.94	0.85	0.89	430
accuracy			0.86	624
macro avg	0.83	0.86	0.84	624
weighted avg	0.87	0.86	0.86	624

MANUAL

	precision	recall	f1-score	support
-1.0	0.76	0.81	0.78	221
1.0	0.89	0.86	0.88	403
accuracy			0.84	624
macro avg	0.83	0.83	0.83	624
weighted avg	0.84	0.84	0.84	624

K Nearest Neighbors (KNN) Classifier:

SKLEARN

	precision	recall	f1-score	support
-1	0.51	0.87	0.64	137
1	0.95	0.76	0.85	487
accuracy			0.79	624
macro avg	0.73	0.82	0.74	624
weighted avg	0.86	0.79	0.80	624

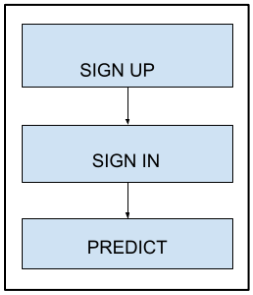
MANUAL

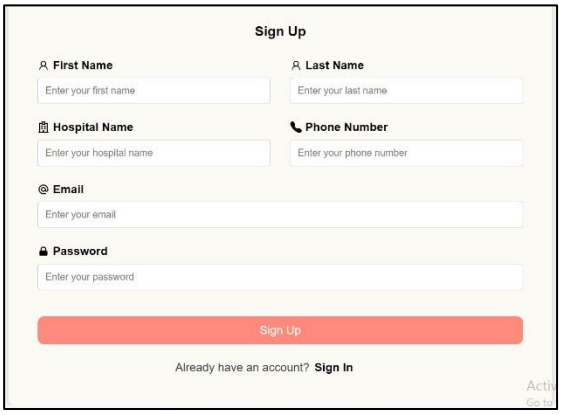
	precision	recall	f1-score	support
-1	0.50	0.88	0.63	132
1	0.96	0.76	0.85	492
accuracy			0.79	624
macro avg	0.73	0.82	0.74	624
weighted avg	0.86	0.79	0.80	624

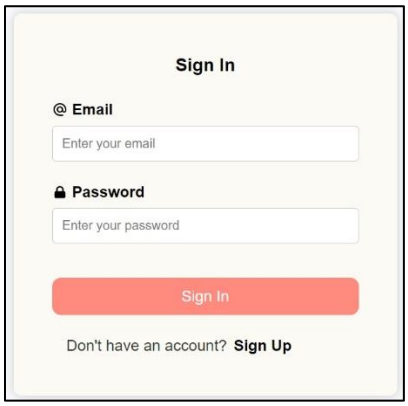
Since the Support Vector Machines algorithm from the sklearn library provided superior results, we saved this model and integrated it into the backend of our software. Additionally, the results of the K Nearest Neighbors (KNN) classifier, both from sklearn and our manual implementation, closely match, demonstrating the robustness of our custom models. Similarly, the SVM results show high consistency between sklearn and custom implementations, highlighting the effectiveness of our approach.

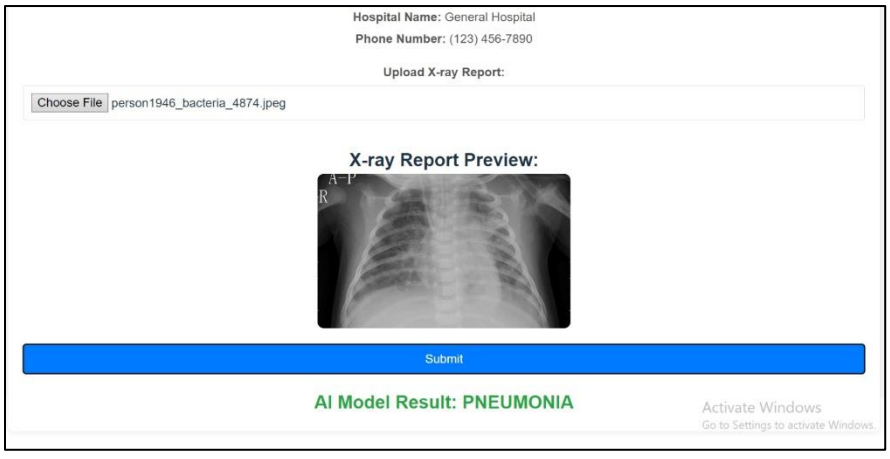
Software Development

Our software is tailored for hospitals and labs, enabling them to predict diagnoses from X-ray images. Users simply register an account, log in, and upload an image to receive accurate predictions swiftly.









CONCLUSION

Since our project focuses on disease prediction, particularly pneumonia, recall is the most crucial evaluation metric. This is because misclassifying an infected person as normal can have serious consequences. Based on our results, SVM has proven to be the most effective model for our use case, ensuring high recall rates and thereby minimizing the risk of false negatives in pneumonia detection.