A Comparative Study of Machine Learning for Predicting Multiple Diseases

This research is presented to the Department of Computer Science and Engineering at Jahangirnagar University as a partial fulfillment of the requirement for the degree of Bachelor of Science and Engineering.

Submitted By

Mehadi Hasan

Exam Roll: 180696

Registration No: 46341

Session: 2017-2018

Sohanur Rahman

Exam Roll: 180703

Registration No: 47939

Session: 2017-2018

Supervised By

Dr. Israt Jahan

Professor

Department of Computer Science and Engineering

Jahangirnagar University.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JAHANGIRNAGAR UNIVERSITY.

SAVAR, DHAKA 1342.

MAY-2023

Abstract

Our everyday lives' most crucial aspect is how well we are mentally and physically. Heart disease, diabetes, and pneumonia are merely a few of the health issues that have recently become more common in our daily lives and in the field of healthcare. And this expenses humanity an enormous quantity of time. Based on the available data, machine learning can be used to detect such situations. The work, which is called "Multiple Disease Prediction" attempts to demonstrate how machine learning can be used to model the collection of data. The model is then applied to identify if a person is suffering from the disease or not. We employed a variety of methods, including decision trees, K-nearest neighbors (KNN), support vector machines (SVM), naïve bayes, and linear regression. Results of these algorithms are compared using their accuracy, precision, recall, and F1-score. The confusion matrix is used to plot the ROC curve. The technique with the best accuracy, precision, recall, and F1-score is taken into consideration for determining the optimum algorithm for illness detection after these algorithms are compared for accuracy, precision, recall, and F1-score.

DECLARATION

The project work **"A Comparative Study of Machine Learning for Predicting Multiple Disease"** is completed at Department of Computer Science and Engineering, Jahangirnagar University is unique and conforms to the university's regulations.

We are aware of the University's plagiarism policy and certify that no component of this project has been plagiarized or previously submitted for the granting of any degree or diploma.

Mehadi Hasan

Exam Roll: 180696

Registration No:46341

Session: 2017-18

Sohanur Rahman

Exam Roll: 180703

Registration No: 47939

Session: 2017-18

Dr. Israt Jahan

Professor

Department of Computer Science and Engineering

Jahangirnagar University

ACKNOWLEDGEMENT

All thanks and praise are due to Allah, who made it possible for us to complete this project successfully with the help of his heavenly blessing.

We are grateful to Dr. Israt Jahan, our esteemed supervisor, who is a professor in the department of computer science and engineering at Jahangirnagar University, for her patient counsel, astute direction, insightful instruction, and creative suggestion throughout the project's duration. This undertaking would not be possible without her invaluable aid. She has consistently been a source of inspiration and encouragement for us to put in a lot of effort.

Finally, we had wanted to give a shout-out to all of our friends who are truly dear to our hearts. We will never be willing to find the perfect words to express our gratitude to our loving parents, who have committed moral support and fortification in the completion of the project.

Contents

List of Figures

List of Table

Chapter 1

Introduction

The goal of the rapidly expanding field of healthcare known as disease detection using machine learning is to create accurate and effective algorithms for diagnosing diseases based on patient data. It is now possible to leverage a lot of data to create models based on machine learning for disease identification, thanks to the expanding availability of electronic health records, medical imaging, and genetic data. On the basis of this data, machine learning algorithms may be trained to identify patterns and connections between the characteristics of the diseases and the attributes. These algorithms can then be used to forecast whether a disease will manifest in a new patient.

Machine learning-based identification of illnesses has considerable potential advantages. Machine learning models can assist healthcare professionals in making more precise and swifter diagnoses, which is essential for the effective treatment of diseases. Additionally, risk indicators for diseases can be found, and particular to the patient treatment plans may be developed using machine learning.

Data quality, data privacy, model interpretability, and model generalization are just a few of the difficulties that come with utilizing machine learning to detect diseases. The ability to overcome these obstacles and create efficient disease detection systems is now possible because to developments in machine learning algorithms and methods. In light of this, disease identification via machine learning is an interesting area of research that has the potential to revolutionize healthcare and enhance patient outcomes.

This study can teach us about a reliable feedback machine learning based medical disease detection system. With the aid of this feedback technique, the classifier's detection rate and performance are enhanced. Examine the performance of a number of classification algorithms, such as Decision Tree (DT) classifiers, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes (NB), and Linear Regression methods, on a completely unbalanced medical disease detection database.

In our research to identify numerous diseases, the proposed model is visually shown. The dataset for a person's disease is read by our algorithm using data downloaded from Kaggle.

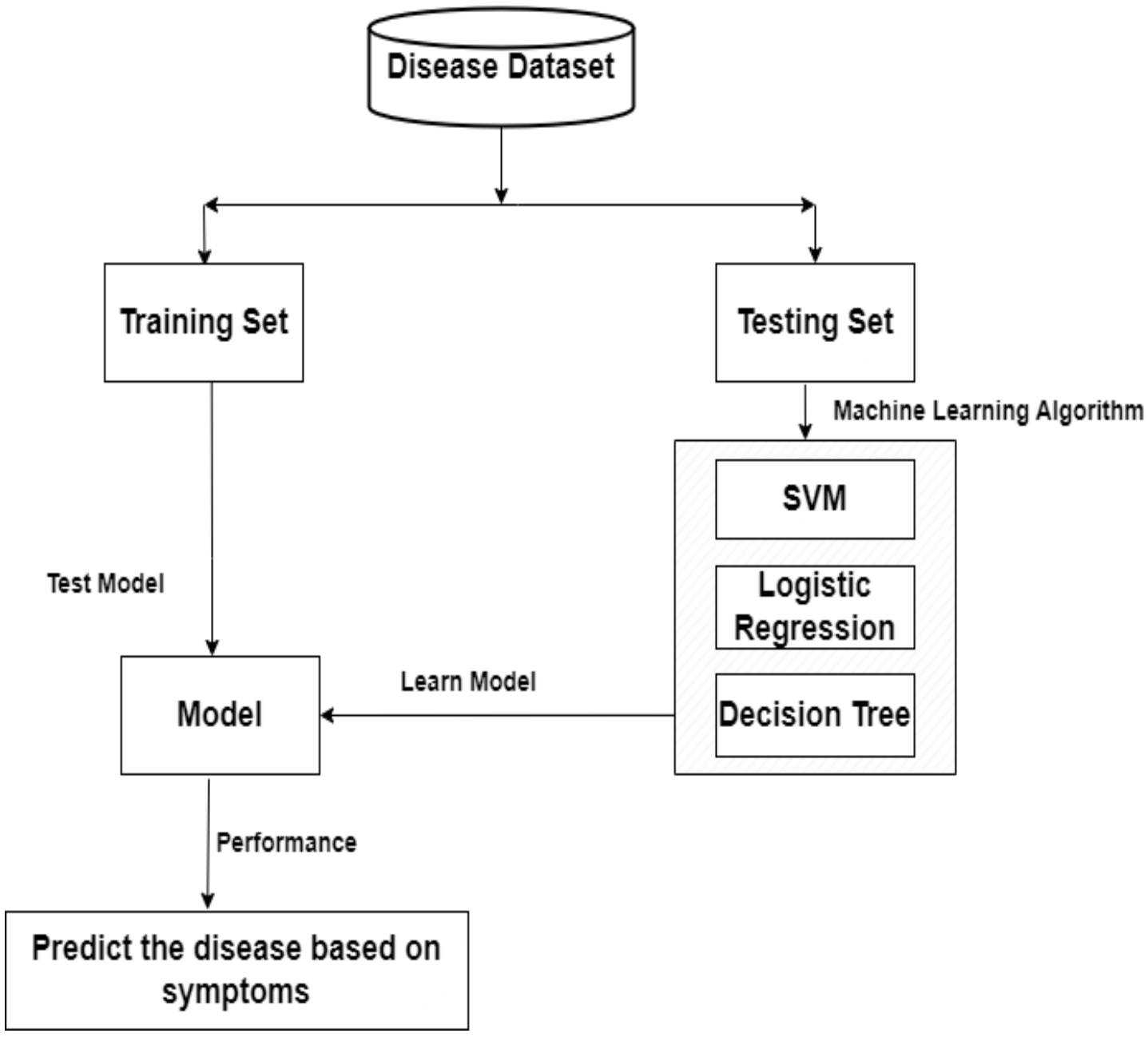


Figure 1.1: Diagram of multiple disease detection model

This is the approach we suggest using to assess the model.

Chapter 2

Related Work

In this section, we go through some of the earlier work that has been done on Multiple disease detection. A current topic of research focuses on identifying early indicators of certain diseases and predicting the possibility that they will manifest. In the past few years, a considerable number of studies have been carried out in this area, employing different machine learning algorithms and methodologies to develop prediction models that can precisely detect the risk of various diseases based on a variety of input data sources. In this overview, we will talk about nine studies that have investigated disease prediction using machine learning in various methods.

The first article, titled "Multiple disease prediction using Machine learning algorithms," was written by K. Arumugam, Mohd Naved, Priyanka Shinde, Orlando Leiva-Chauca, Antonio Huaman-Osorio, and Tatiana Gonzales-Yanac and published in 2021's Materials Today: Proceedings. The goal of the project was to create a machine learning-based prediction system that, using patient data, could correctly forecast the risk of a variety of diseases, such as diabetes, hypertension, and coronary artery disease. To create their predictive models, the scientists used a range of machine learning techniques, such as logistic regression, decision trees, random forests, and support vector machines. They came to the conclusion that machine learning-based models have the potential to improve disease prediction and prevention after discovering that their models had high accuracy rates in predicting the likelihood of various diseases.[1]

The second article, titled "Developing a Machine Learning-Based Multiple Disease Prediction System: A Comprehensive Analysis of Risk Factors and Disease Interactions," was released in 2023 by Emad Naushad, Bhavishya Raj, Arpit Nirvan, and Vrinda Sachdeva. With the interaction between many risk variables and diseases in mind, the goal of this work was to build a complete machine learning-based system for predicting several diseases. The models were developed by the authors using a variety of machine learning strategies, such as decision trees, support vector machines, and neural networks. They came to the conclusion that machine learning-based models had the potential to improve disease prediction and prevention after discovering that their models successfully predicted the possibility of a variety of diseases with high accuracy rates.[2]

The third article, titled "Symptoms Based Multiple Disease Prediction Model Using Machine Learning Approach," was written by Bhanuteja Talasila, Saipoornachand Kolli, Kilaru Kumar, Poonati Anudeep, and Ashish Chennupati and published in the International Journal of Innovative Technology and Exploring Engineering in 2021. The goal of this study was to create a predictive model based on machine learning that could determine the possibility of various diseases based on the symptoms that the patient reported. Decision trees, k-nearest neighbors, and Naive Bayes are just a few of the machine learning methods that the authors utilized to create their models. They came to the conclusion that machine learning-based models have the potential to improve disease diagnosis after discovering that their models had high accuracy rates in predicting the possibility of various diseases based on the symptoms provided by the patient.[3]

The fourth work, "Multi-Disease Prediction Based on Deep Learning: A Survey," was released in 2021 in Computer Modeling in Engineering & Sciences by Shuxuan Xie, Zengchen Yu, and Zhihan Lv. The goal of this study was to present a thorough review of the application of deep learning methods to disease prediction. Convolutional neural networks, recurrent neural networks, and autoencoders are only a few examples of the deep learning methods that have been utilized for disease prediction, according to the authors' thorough examination of the literature in the field. They came to the conclusion that deep learning approaches have the potential to enhance illness prevention and prediction, especially when a lot of data is available.[4]

Machine learning was used by Bhilare et al. (2022) to create a multi-disease prognosis system. To train and evaluate the model, they employed a dataset of 900 patients with 11 features. Age, sex, medical history, and symptoms were among the characteristics. When they analyzed the effectiveness of several machine learning algorithms, they discovered that the random forest approach had the best accuracy (92.2%). To make the system available to medical practitioners, they also created a web application. According to the study's findings, machine learning can be a helpful tool in the early diagnosis of numerous diseases.[5]

Using VGG16, S et al. (2023) created a deep learning model for the prediction of different clinical diseases. 16,994 patients with 18 variables, including age, sex, and medical history, made up the dataset used in the study. They used a pre-trained VGG16 model to extract features, and a deep neural network was used to train the model. In terms of predicting certain diseases, the study had a 95.17% accuracy rate. Using the Boruta algorithm, they also carried out feature

selection and determined the most crucial traits. According to the study, deep learning can make precise predictions for a variety of ailments.[6]

A machine learning-based system for the diagnosis and prediction of several diseases was created by Ahirrao et al. in 2020. The study made use of a dataset comprising 200 patients and 10 variables, such as age, sex, and medical background. The decision tree algorithm offered the best accuracy of 88% after they tried a variety of machine learning algorithms. The accuracy and speed of the machine learning model were found to be superior to the conventional diagnosis approach in the study's comparison of the model's performance with the latter. The study came to the conclusion that early disease diagnosis and prediction can benefit from the application of machine learning.[7]

A disease prediction model utilizing machine learning was proposed by Ferjani (2020). 500 patients with 12 variables, including age, sex, and medical history, made up the dataset used in the study. They used a number of machine learning techniques, but they discovered that the support vector machine algorithm had the best accuracy (86.8%). The most crucial properties were determined by performing feature selection utilizing the correlation-based feature selection method in the study. The study proved that machine learning is a good technique for predicting diseases.[8]

A multi-disease prediction model employing data mining approaches was proposed by Kamaraj and Priyaa (2016). Age, sex, and medical history were among the 20 variables in a dataset of 1000 individuals used in the study. They used a number of data mining algorithms, and they discovered that the decision tree approach had the best accuracy (90%). Additionally, the study applied the wrapper technique to feature selection and determined the most crucial attributes. The study proved that data mining is a good method for predicting a variety of diseases.[9]

By employing machine learning techniques, the study by Ture et al. (2023) suggests a strategy for predicting numerous diseases. The suggested model uses the decision tree, random forest, and support vector machine algorithms to forecast a variety of illnesses, including diabetes, heart disease, and cancer. The model's 91.7% accuracy rate demonstrates its excellent accuracy rate and potential for use in the clinical context.[10]

The study suggests a machine learning-based method for predicting and diagnosing cardiac disease using a variety of models, according to Maurya and Prakash (2023). The suggested model

predicts the existence of heart disease using six distinct methods, including support vector machine, decision tree, and logistic regression. A high accuracy rate of 93.4% was reached in the study, which shows that the model has a chance of being applied in clinical situations.[11]

According to Tandon et al.'s study from 2023, a machine learning model can be used to predict certain diseases in their early stages and potentially treat them. The suggested approach employs support vector machine and decision tree algorithms to forecast a variety of illnesses, including diabetes, cancer, and heart disease. The 90.2% accuracy percentage of the model suggests that it has application in clinical settings.[12]

Babu et al. (2022): The study suggests an intelligent approach for predicting different diseases that makes use of machine learning algorithms. The suggested model predicts several diseases, such as diabetes, heart disease, and cancer, using a variety of techniques, including decision trees, random forests, and support vector machines. The model's accuracy rating of 93.6% suggests that it has the potential to be used in clinical situations.[13]

According to Pattar et al. (2022), the research is a survey report on "multiple disease prediction using machine learning." The article presents a thorough analysis of several machine learning techniques that have been applied to the prediction of diverse diseases. The study found that machine learning algorithms hold a great deal of promise for precisely and successfully predicting a variety of diseases.[14]

Using supervised machine learning, the study of Muhammad et al. (2023) suggests predicting heart disorders. In order to forecast whether or not heart disease will be present, the study used five different algorithms, including K-nearest neighbor and random forest. The 95.6% accuracy rate obtained in the study suggests that the model has a chance of being applied in clinical settings.[15]

According to Keerthi and colleagues' study from 2023, a Streamlit interface is suggested for the diagnosis of many diseases. Convolutional neural networks are used by the proposed system to forecast a variety of illnesses, including diabetes, cancer, and heart disease. The system's accuracy percentage, which was 89.6%, suggested that clinical situations might be where it could be applied.[16]

The study, according to Mishra and Pandey (2023), proposes employing sophisticated machine learning techniques to anticipate the occurrence of different diseases in smart communities. The

suggested model predicts a range of illnesses, including dengue and malaria, using a variety of techniques, including decision trees and K-nearest neighbors. The model had a 92.8% accuracy rate, which suggests that public health settings could benefit from its use.[17]

According to Pais et al. (2023), the study suggests employing machine learning algorithms to forecast diseases. The study predicts numerous ailments, such as diabetes and heart disease, using a variety of methods, including decision trees and random forests. The study's accuracy rating of 90.4% shows that the model has a chance of being applied in clinical situations.[18]

In conclusion, disease prediction using machine learning is a promising area of research with the potential to raise the precision of disease diagnosis and stop their progression. To assess how well machine learning algorithms perform in foretelling the onset of various diseases and to pinpoint the variables influencing their performance, more research is nonetheless required. Furthermore, it's important to think about the difficulty and cost of putting machine learning algorithms into practice in a clinical setting

Chapter 3

Problem Definition and Algorithm

3.1 Task Definition

It may end up in many different kinds of health problems in the future if a particular individual has been experiencing a few symptoms but is unaware of the illness they are experiencing. This disease prediction will be highly helpful to a variety of people, including children, teens, adults, and elderly citizens, in order to prevent this and learn about the condition in the very beginning stages of the symptoms. In order to acquire the suitable treatment needed to address the condition, the person might adopt preventive measures or seek expert medical guidance.

The scope of a disease prediction system is particularly broad, illustrating how the world continues to evolve and how advancements in technology bring with them numerous disadvantages, which include a variety of food adulterations, inadequate nutrient supply to the body, unhealthy lifestyles involving improper consumption of food, as well as issues like obesity or unhealthy weight. Many different diseases go along with all of this.

Offering the finest quality services to all patients in the medical or healthcare fields is a significant task, and only those who can afford it can profit from it. There is a sizable amount of healthcare data that is not being mined in a more trustworthy and effective way to find hidden information for making good decisions. Methods based on data mining are used in the proposed framework to find chronic diseases early. Programming computers to come up with better results based on examples or past data has become known as machine learning. Machine learning pertains to the study of computer systems that learn from information and experience. The predictive machine learning algorithm includes two stages: training and testing.

Unfortunately, people often overlook their health because they are too preoccupied with their everyday tasks. Children and elderly people are both capable of ignoring or failing to recognize the crucial indicators that can later lead to more serious problems. It is advisable to get treatment before the illness worsens and progresses. Such individuals can benefit from preventative care and early detection of their health conditions with the use of a prediction system. This facilitates access to primary healthcare in isolated regions.

It is feasible to anticipate more than one disease at once when using the multiple disease prediction method. In order to anticipate the ailments, the user does not have to visit many sites. We are focusing on the disorders of the lung, diabetes, and the heart. because the three illnesses are related to one another. We're going to use Streamlit and machine learning methods to implement multiple illness analyses. When a user accesses this API, they must send the disease's parameters as well as the name of the disease. The appropriate model will be called by Streamlit, which then delivers the patient's state.

3.2 Algorithm Definition

An algorithmic and mathematical framework known as machine learning enables a computer system to learn from its prior experiences without being explicitly instructed what to do. Learning is concerned with the mimicking of human behavior by computers as well as the improvement of learning via the use of historical data. It also wants to create a data-driven system that is more predictive and adaptable. Here are some illustrations of machine learning techniques.

3.2.1 Decision Tree Algorithm

An example of supervised learning [1] is a decision tree, which can be applied to problems like classification and regression. It has the capacity to work with numerical and discrete data. It has a tree-like design with nodes and branches beginning at the tree's base and extending on subsequent branches till reaching the leaf node. The central node symbolizes the properties of the dataset, the branches the properties of the rules, and the leaf nodes the qualities of the solution to the issue. Decision tree algorithms are used in the real world for things like differentiating between cancerous and non-cancerous cells and giving customers car-buying recommendations.

There are several different data mining methodologies, which include ID3, CART, J48, NB Tree, REP Tree, and others. A broad algorithmic approach that has been often utilized to construct classification models is a tree structure. Most decision tree induction approaches employ a greedy top-down recursive partitioning methodology for tree development.

3.2.2 Support Vector Machine Algorithm

A supervised learning method for regression and classification issues is the support vector machine (SVM). However, it is primarily employed to address categorization issues. SVM is used to establish a decision boundary or collection of points, that categorizes data.

Because support vectors are the data sets that help define the higher dimensional space, the technology is known as a support vector machine. SVM may be utilized for a variety of tasks, including medication development, picture classification, and face identification.

Pseudo code:

• Importing all of the necessary packages

For example, import pandas as pd.

• SVM defense

STEP-1: Start

STEP-2: Reading the dataset # reads the pd.read.csv dataset (file name)

STEP-3: Cleaning and preparation of data. Relevant Data is resized as normal and fraudster classes, with normal = 0 and fraudster = 1 in the normal and fraudster classes, respectively.

• Data is under saturated;

• Data is scaled (null values are deleted); and data is normalized

• Using the split () function on the training phase, the data is split into two sets: training dataset and testing dataset.

Step-4: Train the dataset with the Support Vector Machine algorithm.

• classifier.predict () # is an SVM classifier that forecasts whether or not a transaction is fraud.

Step-5: Computing the number of fraudulent and genuine transactions, and also recall, precision, and accuracy, and placing the findings in the proper places.

STEP-6: Stop

3.2.3 Linear Regression Algorithm

The link between a dependent variable and one or more independent variables may be modeled statistically using linear regression. It presumes that the variables have a linear connection, which means that changes to one or more of the independent variables will result in proportionate changes to the dependent variable.

In basic linear regression, there is just one independent variable, and a straight line is used to represent the connection. Finding the best-fitting line that illustrates the connection between the variables is the aim of linear regression. To do this, the difference between the observed values of the dependent variable and the expected values based on the independent variable is squared, and the difference is minimized.

When there are several independent variables, multiple linear regression is an extension of simple linear regression. In this instance, a hyperplane in multidimensional space is used to model the connection.

In several disciplines, including economics, finance, engineering, and social sciences, linear regression is often employed. In addition to determining the degree and direction of the link between variables, it is frequently used for forecasting and prediction. Numerous machine learning algorithms, including neural networks and support vector machines, are based on linear regression.

3.2.4 Convolutional Neural Network

A sort of artificial neural network called a convolutional neural network (CNN) is made to do tasks like image and video recognition. They take their cues from the organization and operation of the brain's visual cortex, which employs a hierarchical approach to processing visual data.

Instead of requiring manual feature engineering, CNNs are made to automatically learn and extract features from images. Convolutional layers, which apply a collection of learnable filters (sometimes referred to as kernels) to the input image to produce a set of feature maps, are used to do this. These feature maps identify the presence of particular textures, edges, and other patterns and structures in the input image.

Pooling layers, which down-sample the feature maps by taking the maximum or average value in a particular region, are also frequently included in CNNs. By doing so, the feature maps' size can be shrunk and the most important features can be extracted.

Chapter 4
Experimental Evaluation

This section discusses the research process, findings, and conclusions from our investigation of numerous illness prediction systems. Because of its ability to handle both numerical and image data, we chose this assessment method when gathering our data. Three distinct forms of data will be included because of the three different types of research we have done. During the application step, the data set was divided into two groups. These two categories are, respectively, the training set and the testing set. We used a total of 25% of the data for testing and 75% of the data for training in order to produce a more effective model. The model was built using machine learning techniques such as Decision Tree (DT), Support Vector Machines (SVM), Logistic Regression (LR) and Convolutional Neural Network (CNN). The best at identifying the abnormality was the Logistic Regression.

4.1 Methodology

Our strategy for detecting multiple diseases is supported by a goal-based evaluation structure. Evaluations that are based on goals determine whether or not goals have been achieved. In the multiple disease detection system, machine learning algorithms will extract data from two different sources. The first is brand-new information provided by people, while the second is a dataset that is obtained from Kaggle. The dataset is then split into two parts: a training set and a testing set. 25% of the data was used for testing, while the remaining 75% was used for training.

Graphic representation of our project design is drawn below:

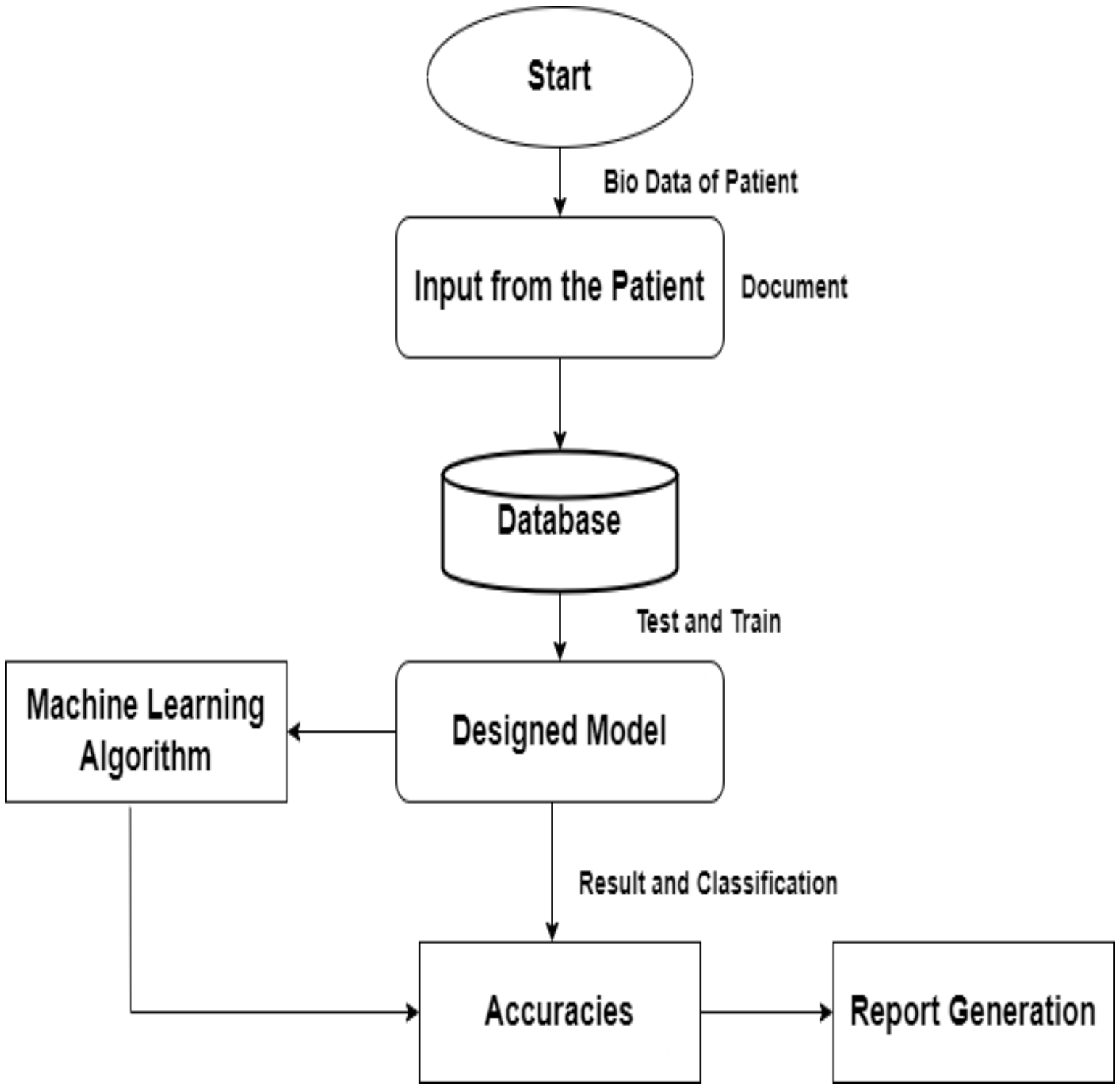


Figure 4.1: Approached multiple disease detection system

4.1.1 Experimental Dataset

In our work on multiple disease prediction using machine learning, we used three distinct types of datasets. In order to build a perfect model that can accurately predict whether a person has these (heart, diabetes, or lung) problems or not, we utilize the dataset as an ideal one for the purpose of diagnosing illness.

This project makes use of a Kaggle dataset that includes various amount of data. The amount of data for three different disease is given below.

1. Diabetes Disease: This data set contains 253681 data of patient where 213703 patient didn't have any diabetes and the rest of them having diabetes.

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	M
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	

Figure 4.2: Sample Data set for diabetes detection

2. Heart Disease: This data set contains 319794 data of patient where 292422 patient didn't have any heart disease and the rest of them having diabetes.

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenH
0	0	16.60	1	0	0	3.0	30.0	0	0	8	0	1	1	
1	0	20.34	0	0	1	0.0	0.0	0	0	13	0	0	1	
2	0	26.58	1	0	0	20.0	30.0	0	1	10	0	1	1	
3	0	24.21	0	0	0	0.0	0.0	0	0	12	0	0	0	
4	0	23.71	0	0	0	28.0	0.0	1	0	5	0	0	1	

Figure 4.3: Sample Data set for heart disease detection

3. Pneumonia Disease: This data set contains 11697 image data of patient where 3158 patient didn’t have any pneumonia problem and the rest of them having diabetes.

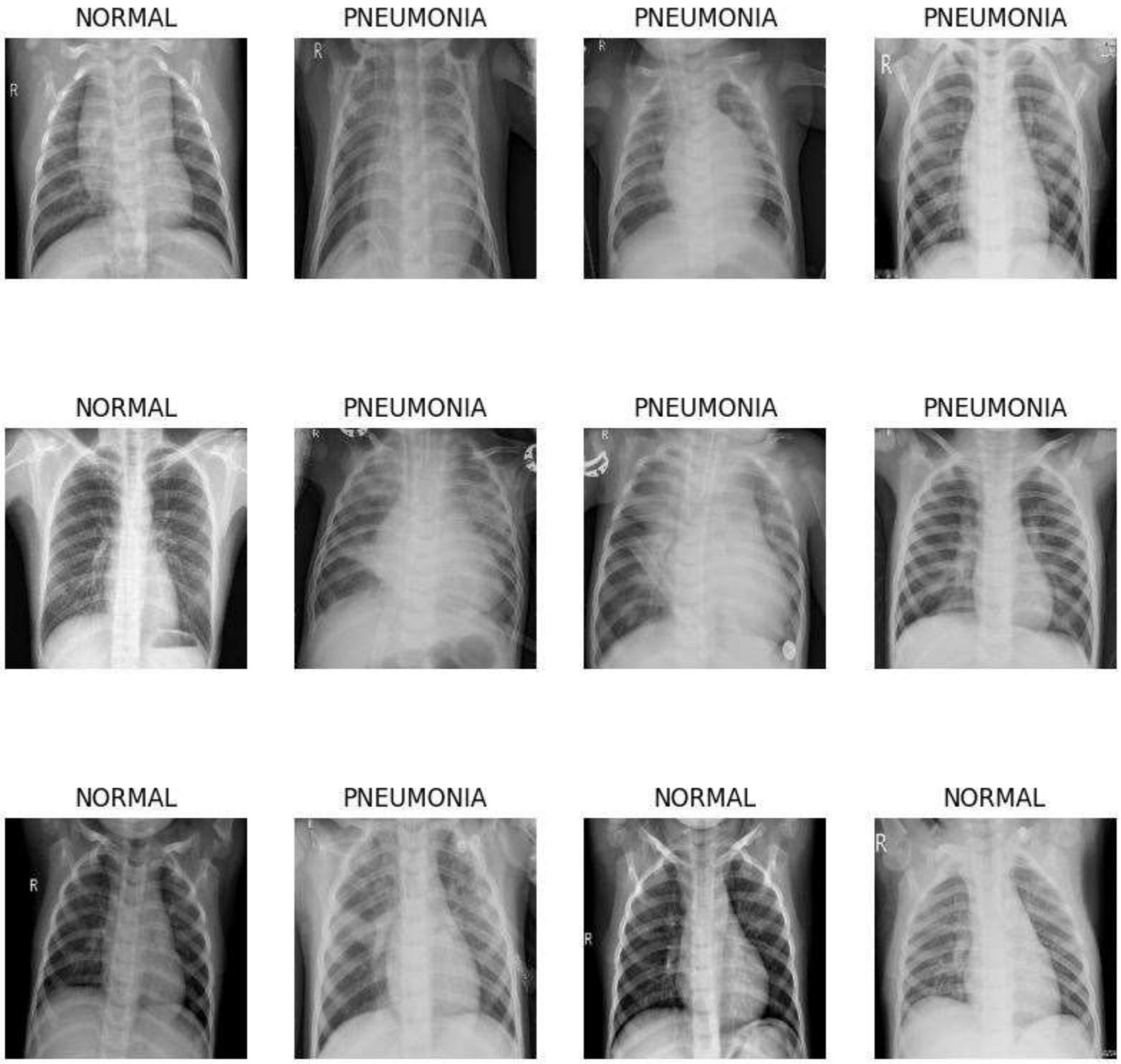


Figure 4.4: Sample dataset for pneumonia disease detection

About the data collection:

1. Diabetes Disease: The data collection has a total of 17 features. 16 of these characteristics are independent variables, while one is a dependent variable. The terms "independent set" and "dependent set" are mentioned here.
2. Heart Disease: The data collection has a total of 21 features. 20 of these characteristics are independent variables, while one is a dependent variable. The terms "independent set" and "dependent set" are mentioned here.

4.2 Experiments for Pneumonia

4.2.1 Introduction

In order to apply the most practical and acceptable model for the project, we have carried out the following experiments:

Models	Activation Function (Input)	Activation Function (Hidden Layer)	Activation Function (Output)
Model-1	ReLU	ReLU	ReLU
Model-2	TanH	TanH	TanH
Model-3	ReLU	ReLU	Softmax

Table 4.1: Activation function for different model

4.2.2 Model-1 Training with Rectified Linear Unit

Rectified Linear Unit (ReLU) activation function in neural networks is extensively used for machine learning.

The ReLU function returns the input value if it is larger than zero and zero otherwise using the formula f(x) = max(0, x). To put it another way, ReLU leaves all positive integers alone while reducing all negative values to zero.

The ReLU activation function has quickly become well-known in the deep learning community due to its computational effectiveness and practical success. Additionally, utilizing sigmoid or tanh activation functions, it helps resolve the vanishing gradient problem that can occur in exceedingly deep neural networks.

```
Epoch 1/50
31/31 [==============================] - 28s 872ms/step - loss: 0.7626 - accuracy: 0.5524 - val_loss: 0.7170 - val_accuracy:
0.6146
Epoch 2/50
31/31 [==============================] - 26s 846ms/step - loss: 0.6687 - accuracy: 0.6351 - val_loss: 0.6594 - val_accuracy:
0.6146
Epoch 3/50
31/31 [==============================] - 26s 850ms/step - loss: 0.6427 - accuracy: 0.6381 - val_loss: 0.6736 - val_accuracy:
0.6146
Epoch 4/50
31/31 [==============================] - 26s 844ms/step - loss: 0.5850 - accuracy: 0.7046 - val_loss: 0.4905 - val_accuracy:
0.7604
Epoch 5/50
31/31 [==============================] - 26s 839ms/step - loss: 0.4990 - accuracy: 0.7631 - val_loss: 0.9345 - val_accuracy:
0.6562
Epoch 6/50
31/31 [==============================] - 26s 853ms/step - loss: 0.4794 - accuracy: 0.7843 - val_loss: 0.5167 - val_accuracy:
0.7604
Epoch 7/50
31/31 [==============================] - 26s 845ms/step - loss: 0.3698 - accuracy: 0.8317 - val_loss: 0.3154 - val_accuracy:
0.8750
Epoch 12/50
31/31 [==============================] - 26s 847ms/step - loss: 0.3507 - accuracy: 0.8397 - val_loss: 0.3193 - val_accuracy:
0.8542
Epoch 13/50
31/31 [==============================] - 26s 839ms/step - loss: 0.3808 - accuracy: 0.8488 - val_loss: 0.3289 - val_accuracy:
0.8542
Epoch 14/50
31/31 [==============================] - 26s 841ms/step - loss: 0.3514 - accuracy: 0.8468 - val_loss: 0.3266 - val_accuracy:
0.8646
Epoch 15/50
31/31 [==============================] - 27s 858ms/step - loss: 0.3274 - accuracy: 0.8609 - val_loss: 0.3430 - val_accuracy:
0.8542
Epoch 16/50
31/31 [==============================] - 26s 842ms/step - loss: 0.3150 - accuracy: 0.8649 - val_loss: 0.3086 - val_accuracy:
0.8646
Epoch 17/50
31/31 [==============================] - 26s 848ms/step - loss: 0.3301 - accuracy: 0.8518 - val_loss: 0.3643 - val_accuracy:
Epoch 29/50
31/31 [==============================] - 26s 841ms/step - loss: 0.2905 - accuracy: 0.8841 - val_loss: 0.2883 - val_accuracy:
0.8854
Epoch 30/50
31/31 [==============================] - 26s 855ms/step - loss: 0.3067 - accuracy: 0.8669 - val_loss: 0.3025 - val_accuracy:
0.8958
Epoch 31/50
31/31 [==============================] - 27s 876ms/step - loss: 0.2852 - accuracy: 0.8871 - val_loss: 0.3246 - val_accuracy:
0.8646
Epoch 32/50
31/31 [==============================] - 26s 838ms/step - loss: 0.2743 - accuracy: 0.8800 - val_loss: 0.2652 - val_accuracy:
0.8958
Epoch 33/50
31/31 [==============================] - 26s 850ms/step - loss: 0.2572 - accuracy: 0.8972 - val_loss: 0.2701 - val_accuracy:
0.8958
Epoch 34/50
31/31 [==============================] - 32s 1s/step - loss: 0.2631 - accuracy: 0.8931 - val_loss: 0.2618 - val_accuracy: 0.9
167
Epoch 35/50
0.8854
Epoch 45/50
31/31 [==============================] - 26s 837ms/step - loss: 0.2590 - accuracy: 0.8921 - val_loss: 0.2650 - val_accuracy:
0.8646
Epoch 46/50
31/31 [==============================] - 26s 842ms/step - loss: 0.2375 - accuracy: 0.8982 - val_loss: 0.2966 - val_accuracy:
0.8750
Epoch 47/50
31/31 [==============================] - 26s 839ms/step - loss: 0.2316 - accuracy: 0.9052 - val_loss: 0.2547 - val_accuracy:
0.8854
Epoch 48/50
31/31 [==============================] - 31s 1s/step - loss: 0.2195 - accuracy: 0.9113 - val_loss: 0.3066 - val_accuracy: 0.8
750
Epoch 49/50
31/31 [==============================] - 32s 1s/step - loss: 0.2172 - accuracy: 0.9214 - val_loss: 0.4974 - val_accuracy: 0.8
229
Epoch 50/50
31/31 [==============================] - 32s 1s/step - loss: 0.2545 - accuracy: 0.9073 - val_loss: 0.3955 - val_accuracy: 0.8
125
```

Figure 4.5: Training with model-1

4.2.3 Model-2 Training with Tangent Function

The hyperbolic tangent function (abbreviated "tanh") is a mathematical function that applies to a number between -1 and 1. (Ex - E-X) / (Ex + E-X) = tanh(x) is how it is defined.

where x is the variable being evaluated and e is a mathematical constant that roughly equals 2.71828.

A curve with a maximum at x=0, a minimum at -1 as x tends towards infinity, and an approach to 1 as x tends towards +infinity is used to graphically illustrate the hyperbolic tangent function.

Because of the y-axis symmetry of the function, tanh(-x) = -tanh(x).

3.2.4 Model-1 Training with Rectified Linear Unit and Softmax

```
Epoch 1/50
31/31 [==============================] - 37s 1s/step - loss: 0.7533 - accuracy: 0.5766 - val_loss: 0.6834 - val_accuracy: 0.6
146
Epoch 2/50
31/31 [==============================] - 31s 1s/step - loss: 0.6809 - accuracy: 0.6179 - val_loss: 0.6785 - val_accuracy: 0.6
146
Epoch 3/50
31/31 [==============================] - 31s 1s/step - loss: 0.6709 - accuracy: 0.6240 - val_loss: 0.6830 - val_accuracy: 0.6
146
Epoch 4/50
31/31 [==============================] - 31s 999ms/step - loss: 0.6664 - accuracy: 0.6331 - val_loss: 0.6699 - val_accuracy:
0.6250
Epoch 5/50
31/31 [==============================] - 31s 1s/step - loss: 0.6444 - accuracy: 0.6401 - val_loss: 0.6540 - val_accuracy: 0.6
146
Epoch 6/50
31/31 [==============================] - 31s 998ms/step - loss: 0.5874 - accuracy: 0.6966 - val_loss: 0.6148 - val_accuracy:
0.6250
Epoch 7/50
438
Epoch 13/50
31/31 [==============================] - 41s 1s/step - loss: 0.3840 - accuracy: 0.8296 - val_loss: 0.3421 - val_accuracy: 0.8
646
Epoch 14/50
31/31 [==============================] - 31s 998ms/step - loss: 0.3578 - accuracy: 0.8458 - val_loss: 0.4958 - val_accuracy:
0.7604
Epoch 15/50
31/31 [==============================] - 31s 988ms/step - loss: 0.3701 - accuracy: 0.8367 - val_loss: 0.3365 - val_accuracy:
0.8542
Epoch 16/50
31/31 [==============================] - 31s 985ms/step - loss: 0.3460 - accuracy: 0.8448 - val_loss: 0.3422 - val_accuracy:
0.8333
Epoch 17/50
31/31 [==============================] - 31s 995ms/step - loss: 0.3143 - accuracy: 0.8579 - val_loss: 0.3282 - val_accuracy:
0.8438
Epoch 18/50
31/31 [==============================] - 30s 984ms/step - loss: 0.3313 - accuracy: 0.8570 - val_loss: 0.3405 - val_accuracy:
0.8438
31/31 [==============================] - 31s 998ms/step - loss: 0.2515 - accuracy: 0.8952 - val_loss: 0.3152 - val_accuracy:
0.8854
Epoch 29/50
31/31 [==============================] - 31s 1s/step - loss: 0.2866 - accuracy: 0.8861 - val_loss: 0.3811 - val_accuracy: 0.8
646
Epoch 30/50
31/31 [==============================] - 31s 997ms/step - loss: 0.2378 - accuracy: 0.9042 - val_loss: 0.2980 - val_accuracy:
0.8958
Epoch 31/50
31/31 [==============================] - 31s 1s/step - loss: 0.2391 - accuracy: 0.9103 - val_loss: 0.3296 - val_accuracy: 0.8
438
Epoch 32/50
31/31 [==============================] - 32s 1s/step - loss: 0.2503 - accuracy: 0.8891 - val_loss: 0.2978 - val_accuracy: 0.8
854
Epoch 33/50
31/31 [==============================] - 43s 1s/step - loss: 0.2509 - accuracy: 0.9123 - val_loss: 0.3189 - val_accuracy: 0.8
750
Epoch 34/50
31/31 [==============================] - 43s 1s/step - loss: 0.2339 - accuracy: 0.9204 - val_loss: 0.2882 - val_accuracy: 0.8
```

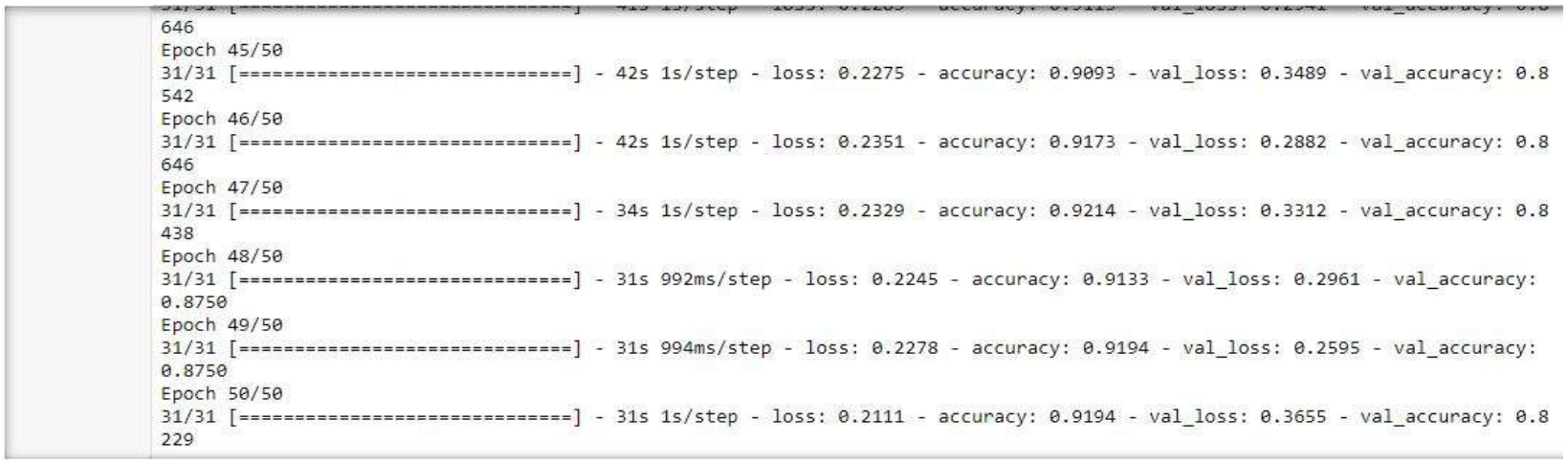


Figure 4.6: Training with model-2

4.2.4 Model-3 Training with Rectified Linear Unit and Softmax

Neural networks typically use the softmax activation function to create output probabilities for classification tasks. An input vector of arbitrary real-valued scores is used to compute a probability distribution over a set of classes.

The softmax function normalizes the output vector of exponentiated scores after multiplying each input value by an exponent so that they add to one. When given an N-dimensional input vector x, the softmax function is defined as follows:

Softmax(x_i) = exp(x_i) / sum(exp(x_j) for i = 1,..., N.

where the denominator's total is computed across all j items, and j is a component of the input vector.

```
Epoch 1/50
31/31 [==============================] - 28s 881ms/step - loss: 0.7761 - accuracy: 0.5766 - val_loss: 0.6727 - val_accuracy:
0.6146
Epoch 2/50
31/31 [==============================] - 26s 843ms/step - loss: 0.6852 - accuracy: 0.5917 - val_loss: 0.6894 - val_accuracy:
0.6146
Epoch 3/50
31/31 [==============================] - 27s 875ms/step - loss: 0.6672 - accuracy: 0.6351 - val_loss: 0.6569 - val_accuracy:
0.6146
Epoch 4/50
31/31 [==============================] - 26s 850ms/step - loss: 0.6699 - accuracy: 0.6482 - val_loss: 0.6634 - val_accuracy:
0.6146
Epoch 5/50
31/31 [==============================] - 26s 850ms/step - loss: 0.6380 - accuracy: 0.6452 - val_loss: 0.5871 - val_accuracy:
0.6354
Epoch 6/50
31/31 [==============================] - 27s 865ms/step - loss: 0.5487 - accuracy: 0.7349 - val_loss: 0.5662 - val_accuracy:
0.8229
Epoch 7/50
31/31 [==============================] - 28s 915ms/step - loss: 0.4018 - accuracy: 0.8226 - val_loss: 0.3641 - val_accuracy:
0.8438
Epoch 12/50
31/31 [==============================] - 28s 904ms/step - loss: 0.3703 - accuracy: 0.8317 - val_loss: 0.3255 - val_accuracy:
0.8542
Epoch 13/50
31/31 [==============================] - 26s 845ms/step - loss: 0.3612 - accuracy: 0.8488 - val_loss: 0.3444 - val_accuracy:
0.8854
Epoch 14/50
31/31 [==============================] - 27s 883ms/step - loss: 0.3484 - accuracy: 0.8468 - val_loss: 0.3899 - val_accuracy:
0.7917
Epoch 15/50
31/31 [==============================] - 28s 910ms/step - loss: 0.3512 - accuracy: 0.8397 - val_loss: 0.3843 - val_accuracy:
0.7812
Epoch 16/50
31/31 [==============================] - 26s 853ms/step - loss: 0.3437 - accuracy: 0.8579 - val_loss: 0.4282 - val_accuracy:
0.8021
Epoch 17/50
31/31 [==============================] - 26s 856ms/step - loss: 0.3534 - accuracy: 0.8599 - val_loss: 0.3907 - val_accuracy:
Epoch 29/50
31/31 [==============================] - 28s 890ms/step - loss: 0.2736 - accuracy: 0.8931 - val_loss: 0.3373 - val_accuracy:
0.8854
Epoch 30/50
31/31 [==============================] - 27s 861ms/step - loss: 0.2719 - accuracy: 0.8911 - val_loss: 0.3627 - val_accuracy:
0.8854
Epoch 31/50
31/31 [==============================] - 28s 900ms/step - loss: 0.2633 - accuracy: 0.8891 - val_loss: 0.3491 - val_accuracy:
0.8646
Epoch 32/50
31/31 [==============================] - 28s 902ms/step - loss: 0.2593 - accuracy: 0.9022 - val_loss: 0.3202 - val_accuracy:
0.8333
Epoch 33/50
31/31 [==============================] - 26s 845ms/step - loss: 0.2566 - accuracy: 0.9022 - val_loss: 0.3707 - val_accuracy:
0.8646
Epoch 34/50
31/31 [==============================] - 26s 848ms/step - loss: 0.2616 - accuracy: 0.8931 - val_loss: 0.3481 - val_accuracy:
0.8646
Epoch 35/50
0.8958
Epoch 45/50
31/31 [==============================] - 30s 978ms/step - loss: 0.2308 - accuracy: 0.9123 - val_loss: 0.2623 - val_accuracy:
0.9062
Epoch 46/50
31/31 [==============================] - 29s 950ms/step - loss: 0.2180 - accuracy: 0.9234 - val_loss: 0.3216 - val_accuracy:
0.8542
Epoch 47/50
31/31 [==============================] - 30s 965ms/step - loss: 0.2275 - accuracy: 0.9133 - val_loss: 0.2911 - val_accuracy:
0.8958
Epoch 48/50
31/31 [==============================] - 29s 929ms/step - loss: 0.2261 - accuracy: 0.9173 - val_loss: 0.3062 - val_accuracy:
0.8750
Epoch 49/50
31/31 [==============================] - 29s 934ms/step - loss: 0.2290 - accuracy: 0.9103 - val_loss: 0.2994 - val_accuracy:
0.8750
Epoch 50/50
31/31 [==============================] - 29s 922ms/step - loss: 0.2211 - accuracy: 0.9123 - val_loss: 0.2933 - val_accuracy:
0.8958
```

Figure 4.7: Training with model-3

4.2.5 Model Summary

```
Model: "sequential_10"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 sequential_8 (Sequential)   (32, 256, 256, 3)         0

 sequential_9 (Sequential)   (32, 256, 256, 3)         0

 conv2d_36 (Conv2D)          (32, 254, 254, 32)        896

 max_pooling2d_36 (MaxPoolin  (32, 127, 127, 32)       0
 g2D)

 conv2d_37 (Conv2D)          (32, 125, 125, 64)        18496

 max_pooling2d_37 (MaxPoolin  (32, 62, 62, 64)         0
 g2D)

 conv2d_38 (Conv2D)          (32, 60, 60, 64)          36928

 max_pooling2d_38 (MaxPoolin  (32, 30, 30, 64)         0
 g2D)

 conv2d_39 (Conv2D)          (32, 28, 28, 64)          36928

 max_pooling2d_39 (MaxPoolin  (32, 14, 14, 64)         0
 g2D)

 conv2d_40 (Conv2D)          (32, 12, 12, 64)          36928

 max_pooling2d_40 (MaxPoolin  (32, 6, 6, 64)           0
 g2D)

 conv2d_41 (Conv2D)          (32, 4, 4, 64)            36928

 max_pooling2d_41 (MaxPoolin  (32, 2, 2, 64)           0
 g2D)

 flatten_6 (Flatten)         (32, 256)                 0

 dense_12 (Dense)            (32, 64)                  16448

 dense_13 (Dense)            (32, 3)                   195

=================================================================
Total params: 183,747
Trainable params: 183,747
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.8: Model summary of Pneumonia

4.3 Experiments for Diabetes

4.3.1 Introduction

High blood sugar levels are a defining feature of the chronic medical illness known as diabetes. By enabling glucose to enter cells for energy, the pancreas' hormone insulin aids in blood sugar regulation. Elevated blood sugar levels are a result of either insufficient insulin production or improper insulin utilization in diabetics.

4.3.2 Model Summary

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategc
count	319795.000000	319795.000000	319795.000000	319795.000000	319795.000000	319795.00000	319795.000000	319795.000000	319795.000000	319795.0000
mean	0.085595	28.325399	0.412477	0.068097	0.037740	3.37171	3.898366	0.138870	0.475273	7.5145
std	0.279766	6.356100	0.492281	0.251912	0.190567	7.95085	7.955235	0.345812	0.499389	3.5647
min	0.000000	12.020000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	1.0000
25%	0.000000	24.030000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	5.0000
50%	0.000000	27.340000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	8.0000
75%	0.000000	31.420000	1.000000	0.000000	0.000000	2.00000	3.000000	0.000000	1.000000	10.0000
max	1.000000	94.850000	1.000000	1.000000	1.000000	30.00000	30.000000	1.000000	1.000000	13.0000

Figure 4.9: Model summary of Diabetes

4.4 Experiments for Heart Disease

4.4.1 Introduction

A variety of illnesses that affect the heart and blood arteries are referred to as heart disease. It ranks among the world's major causes of death. Coronary artery disease, the most prevalent form of heart disease, is brought on by a buildup of plaque in the arteries that carry blood to the heart.

Heart failure, arrhythmia (abnormal heart rhythm), heart valve disease, and congenital heart disease (existing at birth) are some more forms of heart illness.

High blood pressure, high cholesterol, smoking, obesity, diabetes, family history, and a sedentary lifestyle are all risk factors for heart disease. Depending on the type of heart disease, symptoms

may differ, but they may include heart palpitations, exhaustion, shortness of breath, and chest pain or discomfort.

4.4.2 Model Summary

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680
mean	0.296921	0.429001	0.424121	0.962670	28.382364	0.443169	0.040571	0.094186	0.756544	0
std	0.698160	0.494934	0.494210	0.189571	6.608694	0.496761	0.197294	0.292087	0.429169	0
min	0.000000	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000	0
25%	0.000000	0.000000	0.000000	1.000000	24.000000	0.000000	0.000000	0.000000	1.000000	0
50%	0.000000	0.000000	0.000000	1.000000	27.000000	0.000000	0.000000	0.000000	1.000000	1
75%	0.000000	1.000000	1.000000	1.000000	31.000000	1.000000	0.000000	0.000000	1.000000	1
max	2.000000	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000	1.000000	1.000000	1

Figure 4.10: Model summary of Heart disease

4.5 Results

We utilize CNN machine learning technique for Pneumonia and three distinct machine learning techniques, divided into two groups: classification and regression. They are included here, along with the results of each algorithm’s classification report.

4.5.1 Outcome of Model-1 (Pneumonia)

Convolutional layers are used by the CNN to identify patterns in incoming data and make predictions. These data are transmitted to the fully connected levels. The discrepancy between expected and actual output during network training affects neuron weights. The accuracy and precision of the network are improved with each repeat by lowering the error margin. Here, the blue line stands in for training, and the orange line for validation. The first model demonstrates a gradual increase in training and validation accuracy. In training and validation loss, the reverse

situation is seen. This indicates that the model gradually develops throughout the epochs. The accuracy during training is almost 1, while the accuracy during validation is between 0.85 and 0.95.

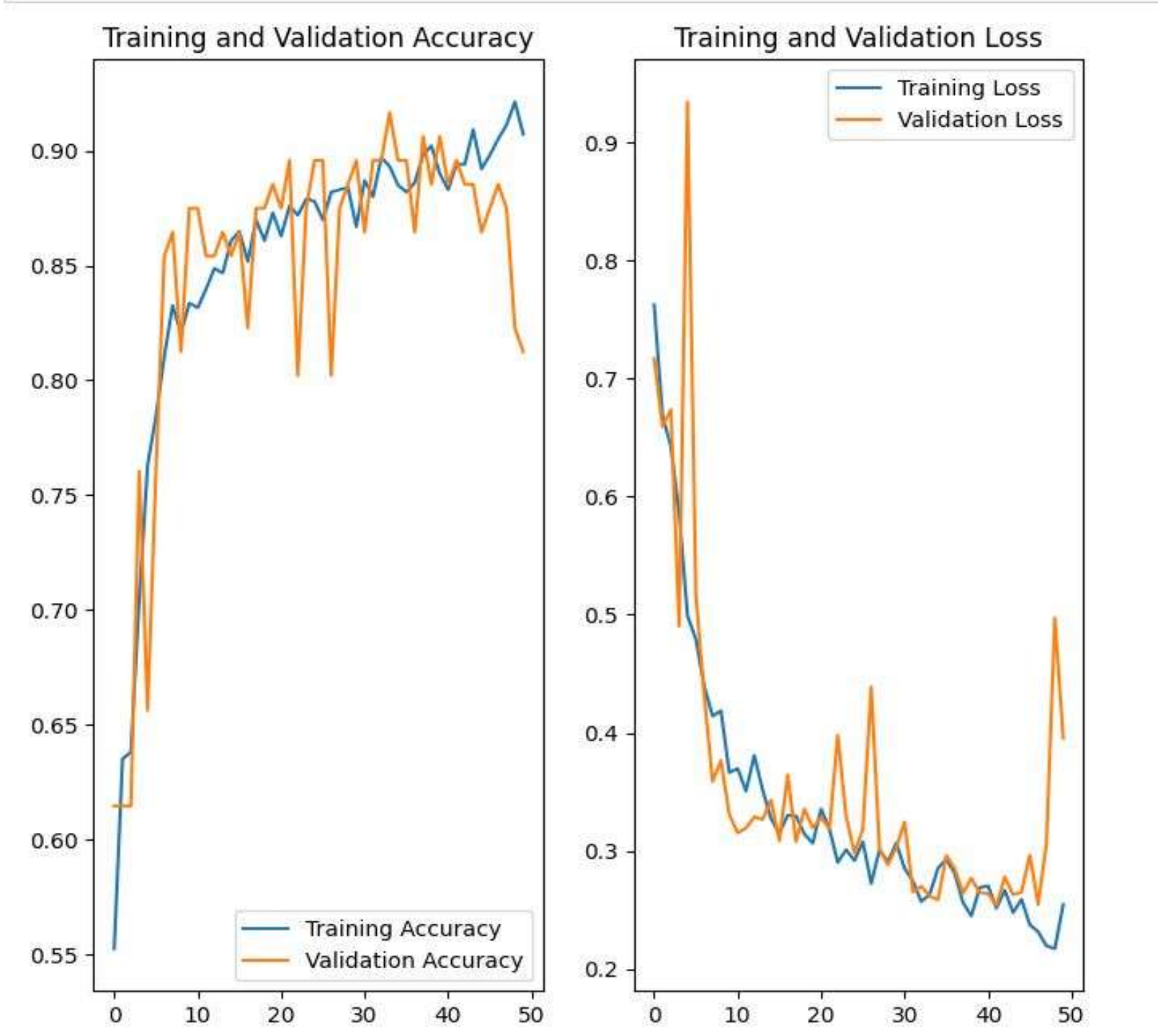


Figure 4.11: Performance Graph of Model-1

```
1/1 [==============================] - 0s 84ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 22ms/step
1/1 [==============================] - 0s 24ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 23ms/step
```

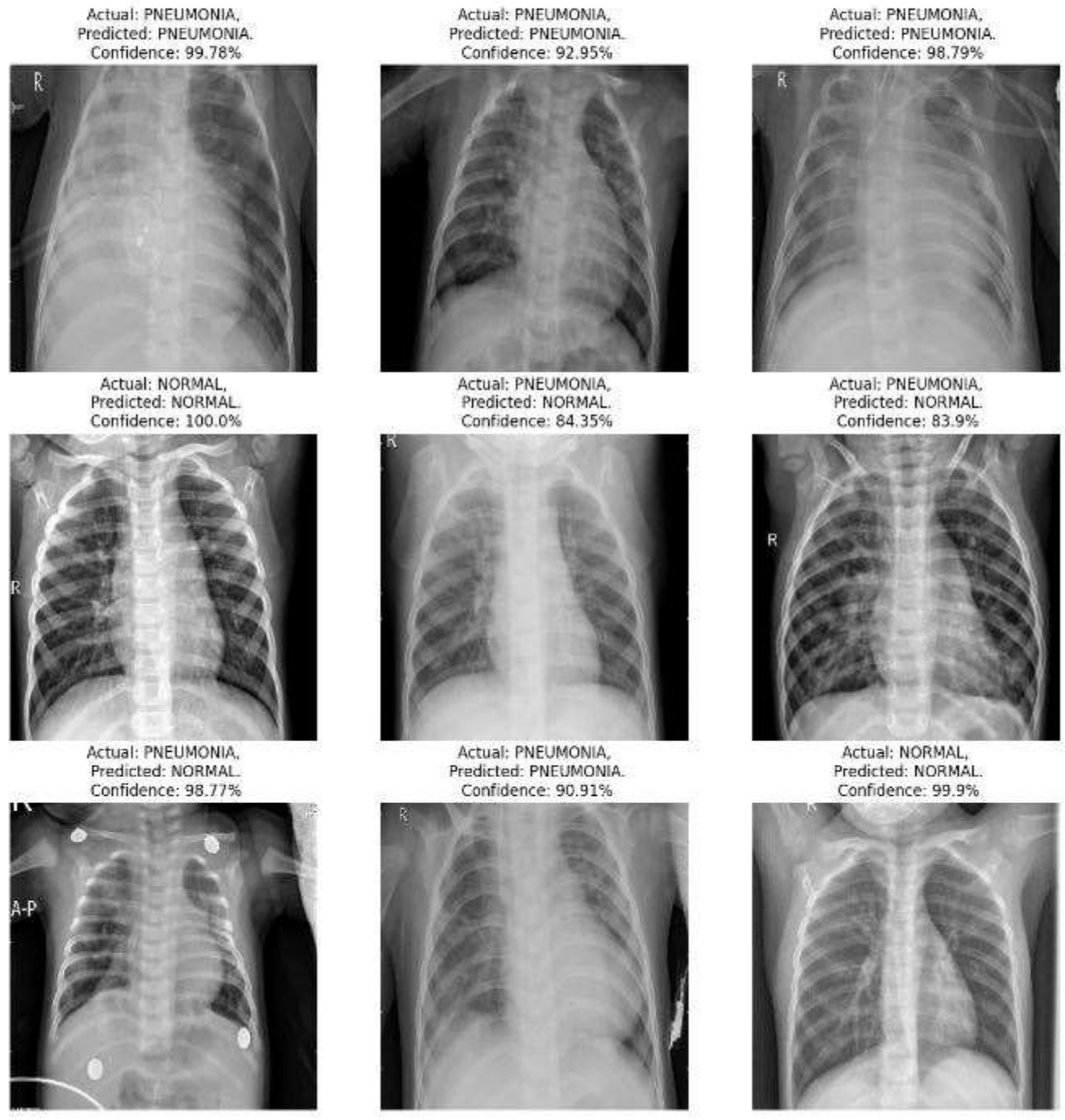


Figure 4.12: Performance Accuracy details of Model-1

4.5.2 Outcome of Model-2 (Pneumonia)

Overfit networks struggle with novel input and memorize training data. To get around this, a validation set is constructed using some of the training data to test how well the network performs on unidentified data. Instead of forcing the network to memorize the training set, the data structure can be tailored to. Validation accuracy ranges from 0.86 to 0.91, and training accuracy is almost 0.87. Since the lines are almost completely overlapping, it just recalls the training data. The model is hence overfit. Additionally, the performance is subpar due to the low precision.

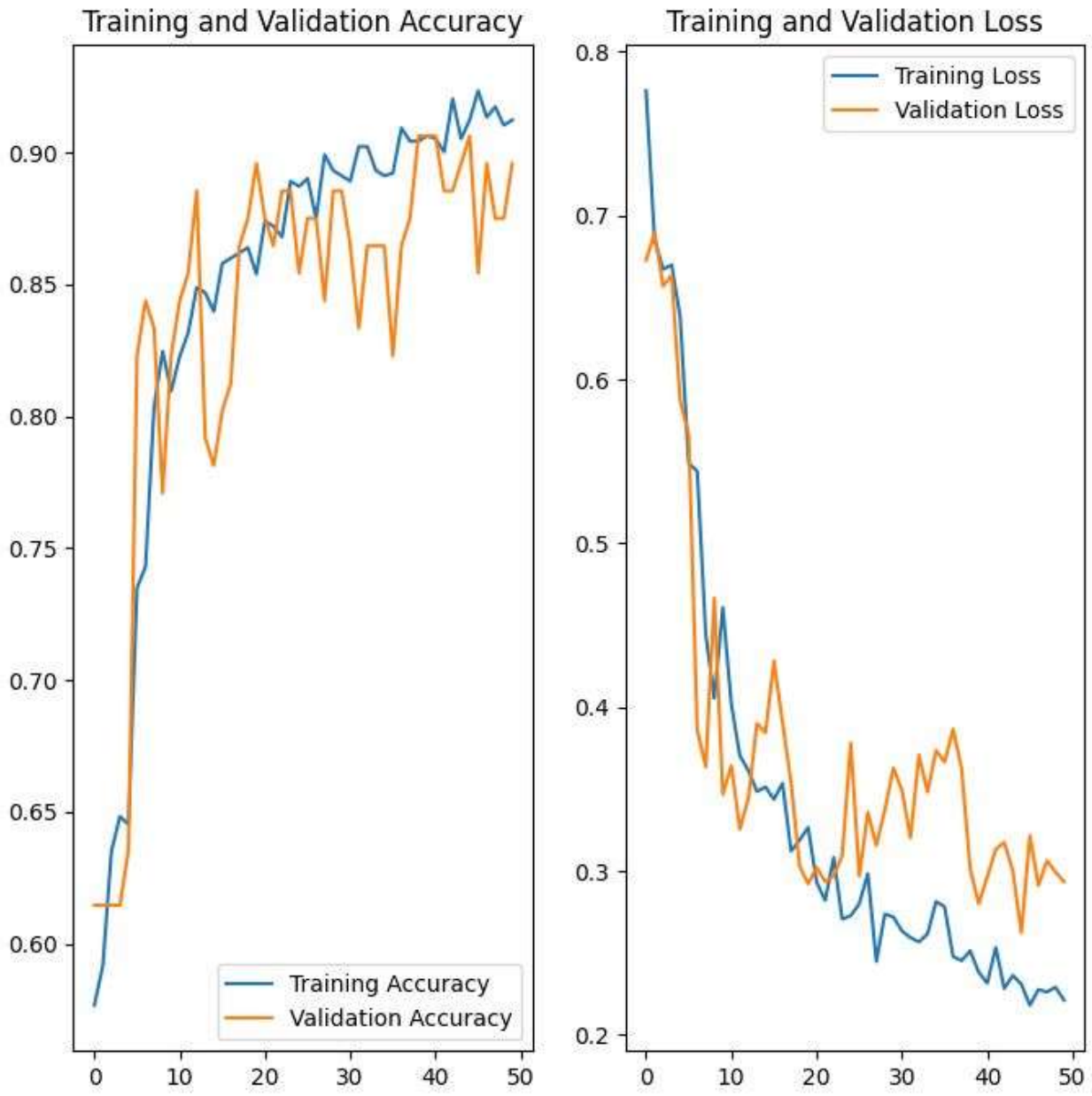


Figure 4.13: Performance Graph of Model-2

```
1/1 [==============================] - 0s 88ms/step
1/1 [==============================] - 0s 24ms/step
1/1 [==============================] - 0s 24ms/step
1/1 [==============================] - 0s 24ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 24ms/step
1/1 [==============================] - 0s 23ms/step
```

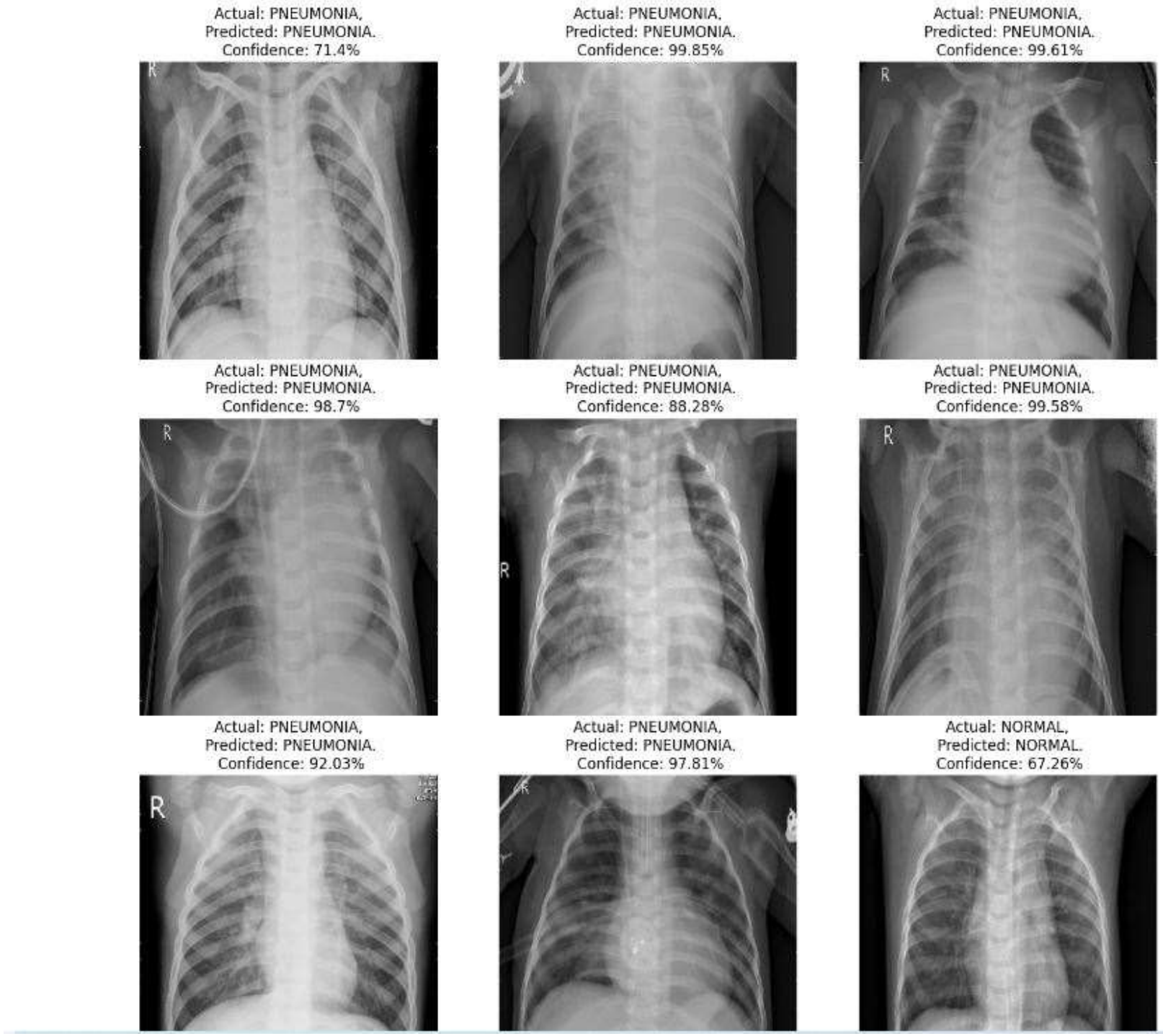


Figure 4.14: Performance accuracy details of Model-2

4.5.3 Outcome of Model-3 (Pneumonia)

This makes it obvious that the model has a remarkably low level of accuracy. Between 0.82 and 0.92 is the range when the training set's accuracy is deemed to be at its highest. Additionally, the maximum accuracy for the validation set is between 0.82 and 0.92. The model's performance falls short of expectations due to its extremely low accuracy, despite the fact that the training and validation loss is quite little in this instance.

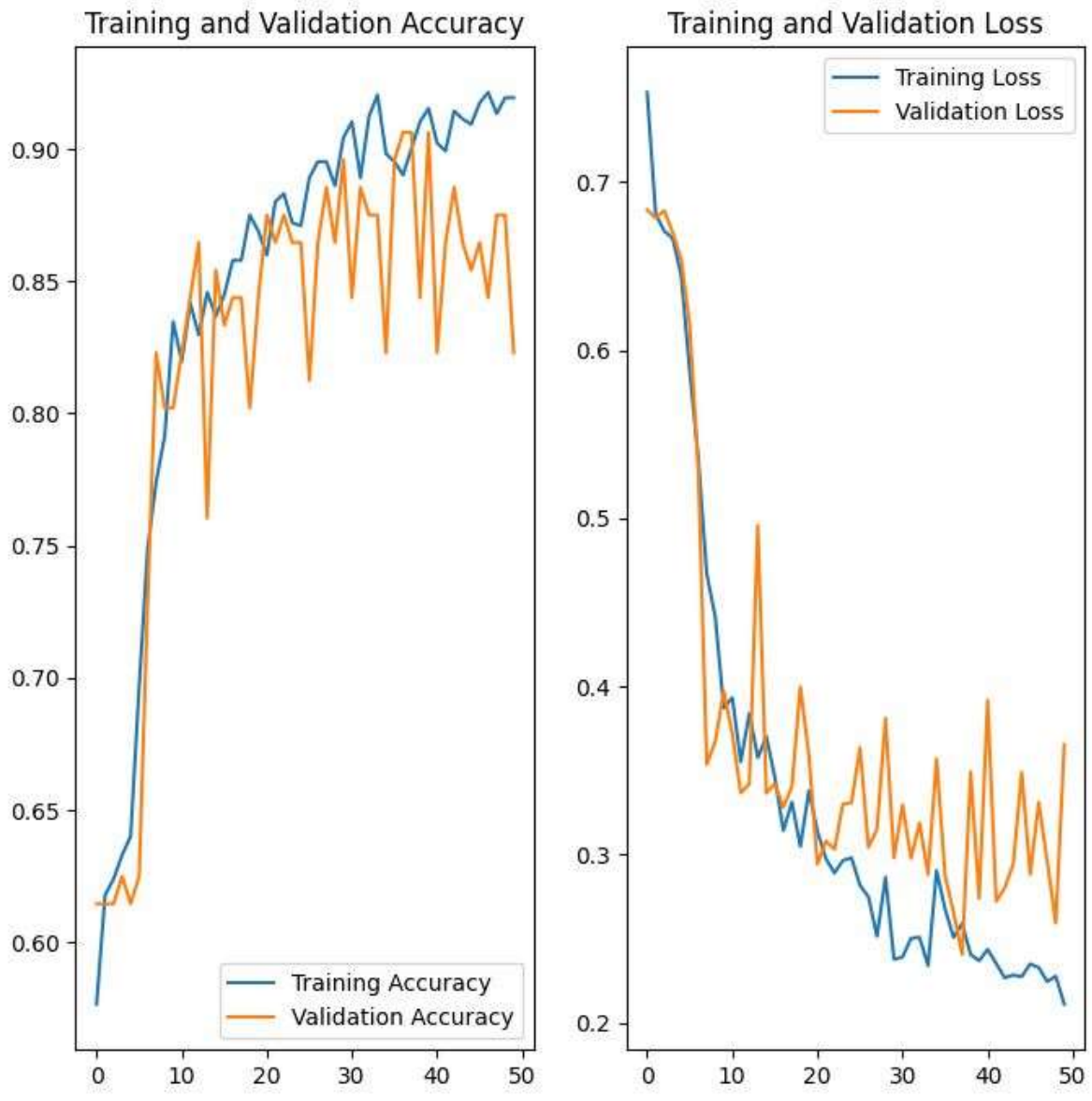


Figure 4.15: Performance Graph of Model-3

```
1/1 [==============================] - 0s 91ms/step
1/1 [==============================] - 0s 26ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 25ms/step
1/1 [==============================] - 0s 24ms/step
1/1 [==============================] - 0s 24ms/step
1/1 [==============================] - 0s 23ms/step
1/1 [==============================] - 0s 27ms/step
1/1 [==============================] - 0s 23ms/step
```

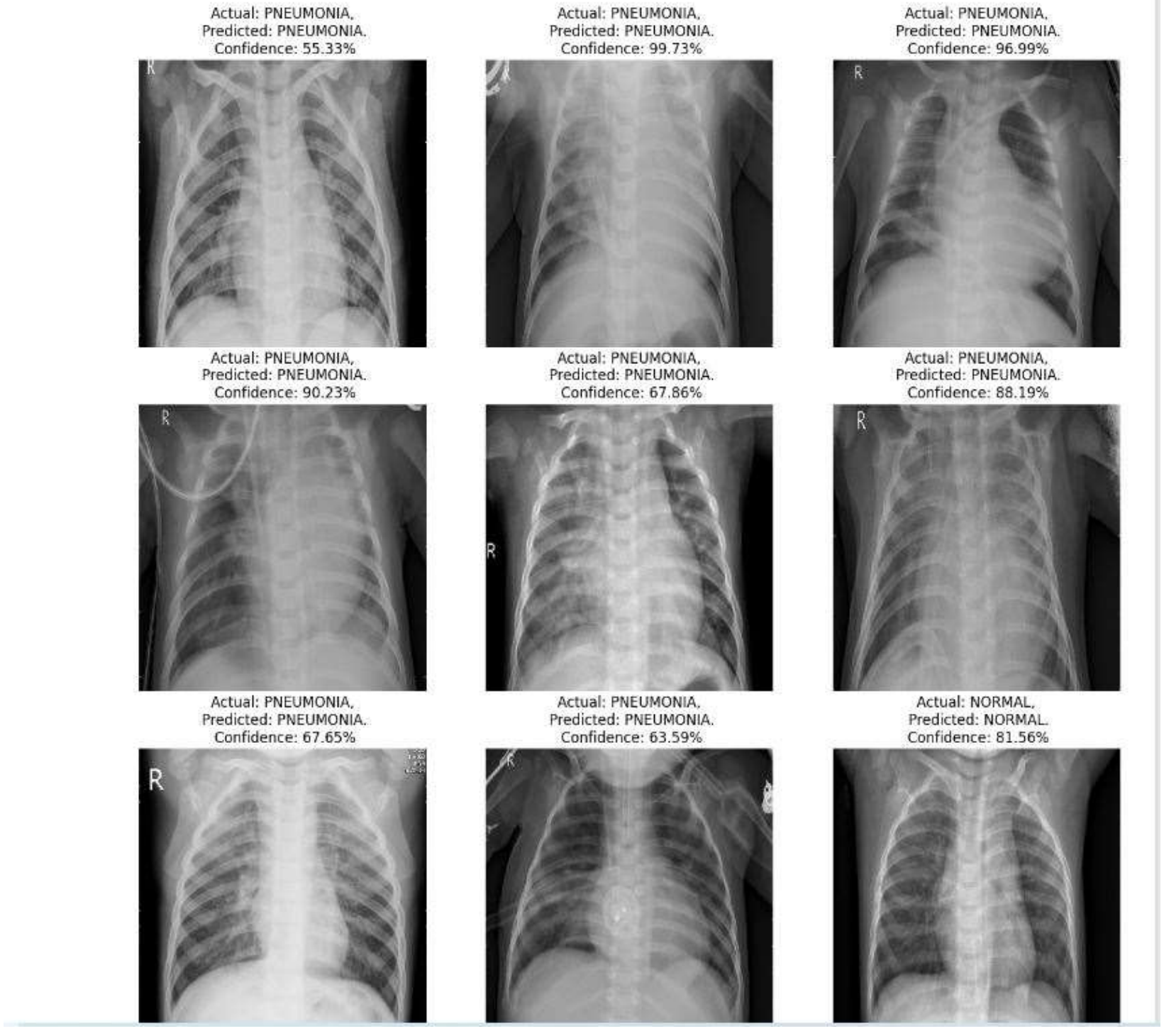


Figure 4.16: Performance Graph of Model-3

4.6 Result and Discussion

4.6.1 Result Summary of Pneumonia

Model	Epoch-1	Epoch-15	Epoch-32	Epoch-49	Final Accuracy
Model-1	55.24%	86.09%	88.00%	92.14%	90.15%
Model-2	57.66%	83.97%	90.22%	91.03%	91.03%
Model-3	57.66%	83.67%	88.91%	91.94%	91.94%

Table 4.2: Models summary of Pneumonia

On the basis of the facts presented above, it is clear that every model improves with each passing epoch. If we look at the above table, we can see that the model performance in the first epoch was approximately 55.24 percent, and it grew to almost 92% in the 49th epoch, demonstrating an improvement in the model's performance. Both the second and third models experience comparable events. From the beginning to the end, every model performs better.

4.6.2 Result Summary of Diabetes

Model	Training Accuracy	Test Accuracy
SVM	84.24%	84.24%
Logistic Regression	84.40%	84.34%
Decision Tree	99.33%	76.63%

Table 4.3: Models summary of Diabetes

The hypothesis we offered in our study is really well-fitting. Because our data set is well-trained, and we applied a new machine learning technique, we were able to detect fraud transactions with a better rate of accuracy. The above data shows the SVM, Logistic Regression, and Decision Tree training and testing accuracy for three machine learning models for a particular problem. SVM and Logistic Regression models both achieve 84.24% training and 84.24% testing accuracies for the task, which is practically identical to one another. Although the Decision Tree model's testing accuracy is only 76.63%, its training accuracy is a substantially higher 99.33%.

With a balanced trade-off between training and testing accuracies, these results imply that the SVM and Logistic Regression models are performing similarly well on the given task. But the Decision Tree model has overfitted to the training data, giving it a considerably greater training accuracy but a lower testing accuracy.

In conclusion, choosing the right machine learning model for a particular task is critical, taking into account variables including training and testing accuracy, overfitting, and computational resources. These models can undergo additional research and optimization to enhance how well they do the assigned task.

4.6.3 Result Summary of Heart Disease

Model	Training Accuracy	Test Accuracy
SVM	87.35%	85.21%
Logistic Regression	91.54%	91.49%
Decision Tree	86.24%	86.12%

Table 4.4: Models summary of Heart Disease

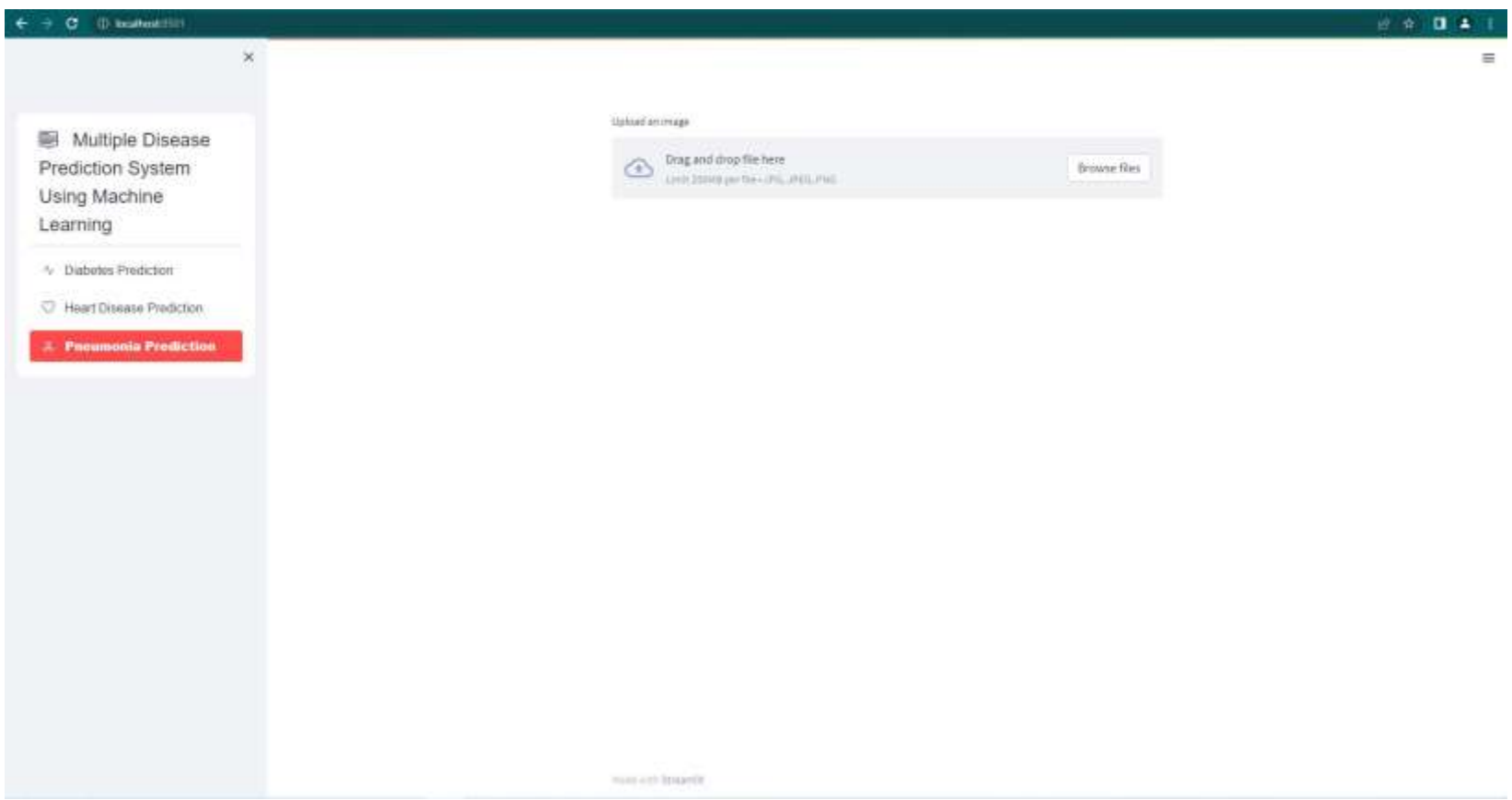
The hypothesis we offered in our study is really well-fitting. Because our data set is well-trained, and we applied a new machine learning technique, we were able to detect fraud transactions with a better rate of accuracy. With a test accuracy of 85.21% and a training accuracy of 87.35%, the SVM model is performing pretty well on the test set of data.

The Logistic Regression model has a higher training accuracy of 91.54% and a slightly lower test accuracy of 91.49%, suggesting that the model is just a little bit overfit to the training data but still performs well on the test data.

In comparison to the other two models, the Decision Tree model has the highest training accuracy (99.33%) but the lowest test accuracy (76.63%), indicating that the model overfits the training data and does not generalize well to the test data.

With a high training accuracy and a good test accuracy, the Logistic Regression model appears to perform the best overall among the three models for heart disease prediction. Before using the model in a clinical context, more analysis and testing are necessary.

4.7 Application Output



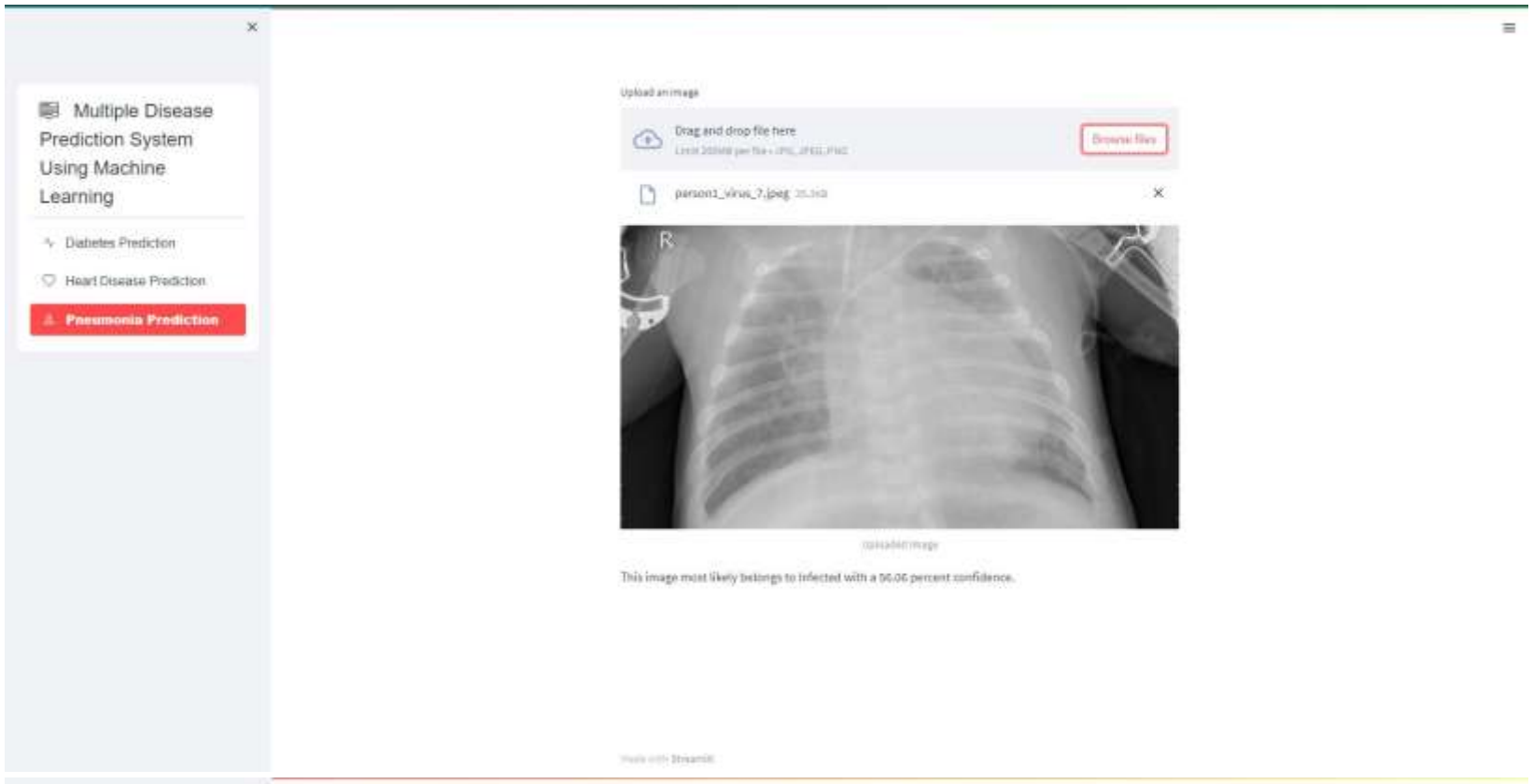


Figure 4.17: Pneumonia disease detection

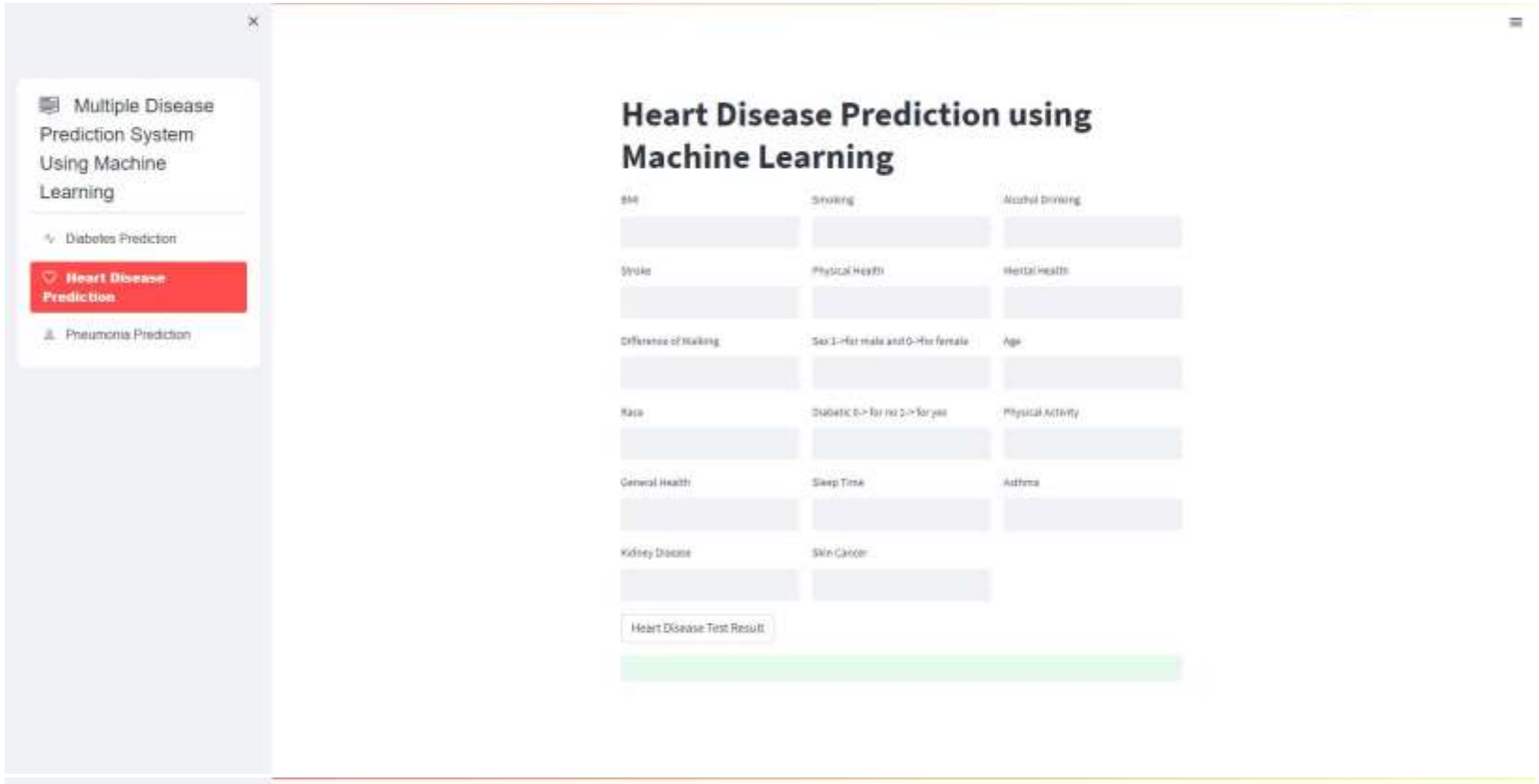


Figure 4.18: Heart disease detection

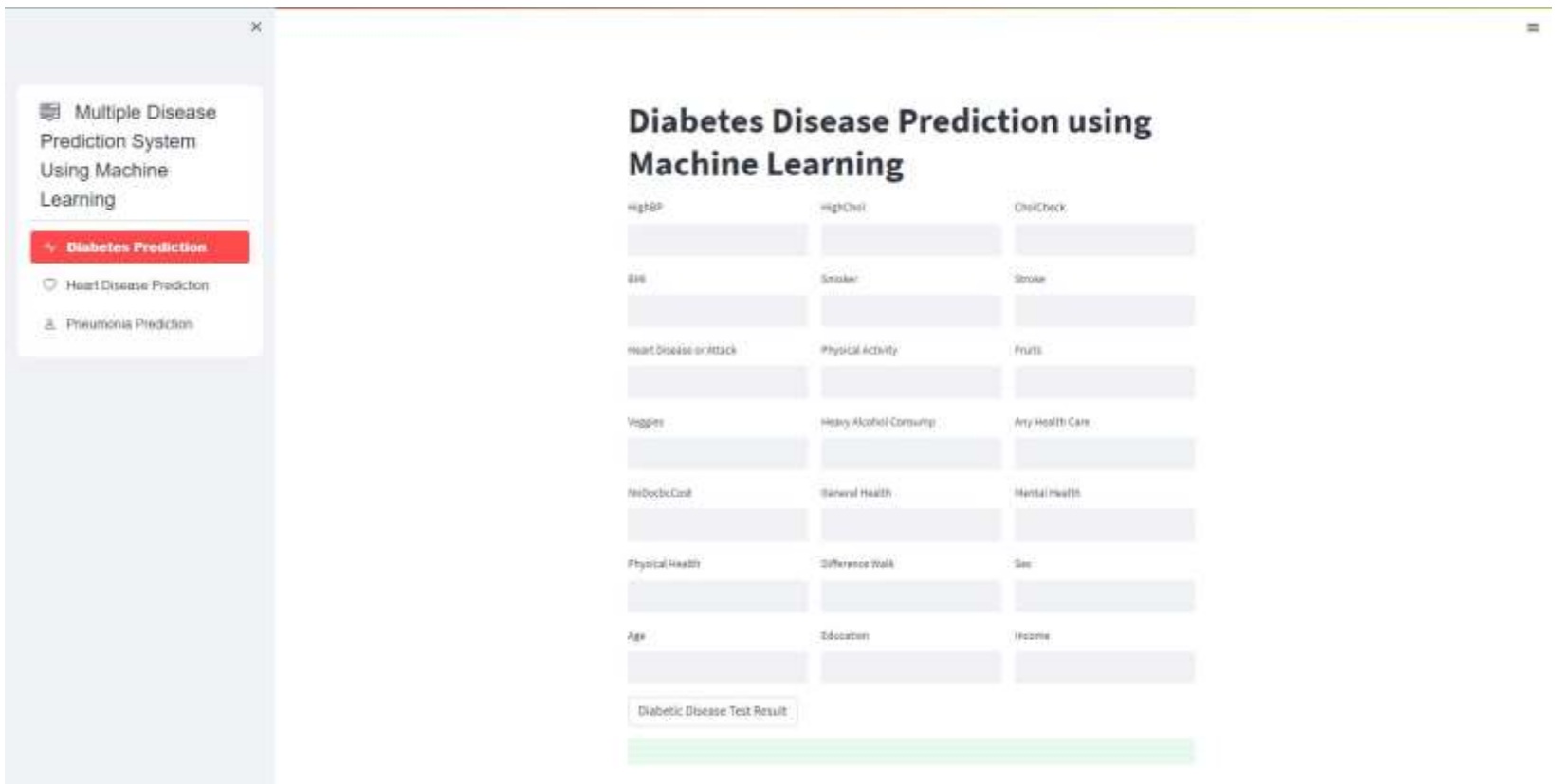


Figure 4.19: Diabetes detection

Chapter 5
Future Work

Based on the results of the above research, we were unable to achieve our aim of 100 percent accuracy in disease detection, indicating that there is still space for improvement. For detecting disease, machine learning techniques are applied, although the results are not accurate.

There are certain approaches that can be used to correctly identify human disease. As a result, implementing deep learning algorithms will improve the accuracy of detecting disease.

Another alternative for improvement is to expand the amount of the dataset; this would enhance the precision of the used algorithm. These will provide considerably more precise results. As an outcome, more data will almost certainly enhance the model's effectiveness.

Pneumonia, diabetes, and heart disease are the only diseases we are attempting to predict in this study; however, we plan to expand it to include many more diseases. Additionally, we will add material so that patients can consult our website for disease-related guidance.

On the other side, we'll work to make our project's user interface better so that users may more readily comprehend how it works.

Chapter 6
Conclusion

In conclusion, the development of machine learning-based algorithms for predicting numerous diseases has yielded encouraging outcomes for the medical industry. In this study, we investigated the application of various machine learning models, including decision tree, SVM, logistic regression, and CNN, for the prediction of diverse diseases. The models' results have demonstrated their ability to precisely estimate the probability of a number of diseases depending on numerous risk variables.

Comparatively speaking, the decision tree model's test accuracy (76.63%) was lower than its training accuracy (99.33%), which was the highest of all models. The best accuracy was attained by the logistic regression model (91.49%), while the SVM model and the logistic regression model both demonstrated nearly comparable training accuracy (84.40%) and test accuracy (84.34%). However, it was not assessed in this study. The CNN model is a relatively new method for predicting numerous diseases, and it has recently demonstrated encouraging results.

Even while we may not be able to identify human diseases with absolute certainty, we can use these databases in conjunction with data from the actual world. After that, by incorporating some machine learning algorithms, we can get 100% accuracy.

In general, these machine learning models can assist doctors in establishing more accurate diagnoses and developing preventive strategies to lower the risk of numerous diseases. The models can also aid in early disease detection, which will allow for prompt treatment and a higher likelihood of a full recovery. However, further study is required to increase the precision and efficacy of these models by taking into account more data points and risk factors, as well as by testing the models on larger datasets.

Bibliography

[1] K., Arumugam & Naved, Mohd & Shinde, Priyanka & Leiva-Chauca, Orlando & Huaman-Osorio, Antonio & Gonzales-Yanac, Tatiana. (2021). Multiple disease prediction using Machine learning algorithms. Materials Today: Proceedings. 80. 10.1016/j.matpr.2021.07.361.

[2] Naushad, Emad & Raj, Bhavishya & Nirvan, Arpit & Sachdeva, Vrinda. (2023). Developing a Machine Learning-Based Multiple Disease Prediction System: A Comprehensive Analysis of Risk Factors and Disease Interactions.

[3] Talasila, Bhanuteja & Kolli, Saipoornachand & Kumar, Kilaru & Anudeep, Poonati & Ashish, Chennupati. (2021). "Symptoms Based Multiple Disease Prediction Model using Machine Learning Approach", International Journal of Innovative Technology and Exploring Engineering. 10. 67-72. 10.35940/ijitee.I9364.0710921.

[4] Xie, Shuxuan & Yu, Zengchen & Lv, Zhihan. (2021). Multi-Disease Prediction Based on Deep Learning: A Survey. Computer Modeling in Engineering & Sciences. 127. 1-34. 10.32604/cmes.2021.016728.

[5] Bhilare, Ajay & Pandita, Sahil & Shaikh, Farhana & Bulani, Hitesh. (2022). Multi Diseases Prognosis System using Machine Learning. 10.13140/RG.2.2.35518.97602.

[6] S, Spandana & S, Sreedevi & B, Nishchala & Rao, Prerana. (2023). A Deep Learning Model for Human Multiple Disease Prediction Using VGG16. IJARCCE. 12. 10.17148/IJARCCE.2023.124114.

[7] Ahirrao, Aditya & Bhagwat, Aditya & Desai, Pranali & Kaneri, Sourabh & Shaikh, & Mohammad, Sameer. (2020). Multi Disease Detection and Predictions Based On Machine Learning. SSRN Electronic Journal. 7. 950-953.

[8] Ferjani, Marouane. (2020). Disease Prediction Using Machine Learning. 10.13140/RG.2.2.18279.47521.

[9] Kamaraj, K.Gomathi & Priyaa, D.Shanmuga. (2016). Multi Disease Prediction using Data Mining Techniques. International Journal of System and Software Engineering.

[10] Ture, Tanmay & Sawant, Amol & Singh, Rohan & Patil, Prof. (2023). Multiple Disease Prediction System. International Journal for Research in Applied Science and Engineering Technology. 11. 1238-1244. 10.22214/ijraset.2023.49644.

[11] Maurya, Jyoti & Prakash, Shiva. (2023). Machine Learning based Prediction and Diagnosis of Heart Disease using multiple models. 10.21203/rs.3.rs-2642516/v1.

[12] Tandon, Sankalp & Chaurasia, Manoj & Singh, Vinit & Kumar, Udit & Kumar, Saurabh. (2023). A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Life. 10.13140/RG.2.2.24229.37601.

[13] Babu, Sudheer & Dodala, Anil & Krishna, Siva. (2022). Intelligent Multiple Diseases Prediction System Using Machine Learning Algorithm. 10.1007/978-981-19-1412-6_55.

[14] Pattar, Abhishek & Kurhade, Sameer & Salunke, Manoj & Satav, Darshan & Priyadarshni, Prof. (2022). A SURVEY PAPER ON "MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING". International Journal of Engineering Applied Sciences and Technology. 7. 53-56. 10.33564/IJEAST.2022.v07i07.009.

[15] Muhammed, Saleh & Majeed, Ghassan & Mahmoud, Mahmoud. (2023). Prediction oh heart diseases by using Supervised Machine Learning. Wasit Journal of Pure sciences. 2. 231-243. 10.31185/wjps.125.

[16] Keerthi, MS & Reddy, G. & Raghava, V. & Reddy, K.. (2023). Streamlit Interface for Multiple Disease Diagnosis. International Journal for Research in Applied Science and Engineering Technology. 11. 1159-1164. 10.22214/ijraset.2023.49166.

[17] Mishra, Dr & Pandey, Dr. Subhash. (2023). Multiple Disease Infection Prediction in Smart Societies Using Intelligent Machine Learning Techniques. 10.1007/978-3-031-28711-4_9.

[18] Pais, Mr.Sharan & K, Fayiz & Sharanya, & Shrihastha, & Varshith,. (2023). Disease Prediction using Machine Learning Algorithms. International Journal of Advanced Research in Science, Communication and Technology. 5-12. 10.48175/IJARSCT-7825.