Multi Disease Prediction System

Divya Mandem[1], B. Prajna[2]

[1]*PG Scholar, Dept. of computer science and System Engineering(A), Andhra University College of Engineering Vishakhapatnam, Andhra Pradesh*

[2]*Professor, Dept. of computer science and System Engineering(A), Andhra University College of Engineering Vishakhapatnam, Andhra Pradesh*

***Abstract* - Multi Disease Prediction" system based on predictive modeling predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user as input and gives the probability of the disease as an output Disease Prediction is done by implementing the random forest classifier and also, we implement Malaria and Pneumonia using Deep Learning Model(CNN) and in the prosed method it gives the better accuracy and we design web allocation for prediction system.**

***Index Terms* - Random Forest, Chronic Disease, CNN, Predictive Modeling.**

I.INTRODUCTION

When anyone is currently afflicted with an illness, they must see a doctor, which is both time consuming and expensive. It can also be difficult for the user if they are out of reach of doctors and hospitals because the illness cannot be detected. So, if the above procedure can be done using an automated software that saves time and money, it could be better for the patient, making the process go more smoothly. There are other Heart Disease Prediction Systems that use data mining methods to analyses the patient's risk level. Disease Predictor is a web-based program that predicts a user's disease based on the symptoms they have. Data sets from various health-related websites have been obtained for the Disease Prediction system. The consumer will be able to determine the likelihood of a disease based on the symptoms given using Disease Predictor. People are always curious to learn new things, particularly as the use of the internet grows every day. When an issue occurs, people often want to look it up on the internet. Hospitals and physicians have less access to the internet than the general public. When people are afflicted with an illness, they do not have many options. As a result, this system can be beneficial to people. Chronic illness is a disease that lasts a long time or takes a long time to heal, and many chronic diseases cannot be cured but can only be managed with daily treatments. India, like all other nations, is undergoing significant social and economic shifts, which is causing a rapid rise in the prevalence of cardiovascular disease. Many developed, developing, and developing countries, including India, are dealing with a wide range of chronic diseases, especially cardiovascular disease and diabetes, which could have serious consequences for global health, security, and economy. The rapid urbanisation and economic growth of today's world has resulted in a wide range of lifestyles. Chronic diseases are now a problem in all nations, with chronic disease afflicting one-third of the population in each. Chronic disease care is more expensive, and it is difficult for those who are sick. In the medical field, a large number of chronic disease datasets are gathered and processed, and data mining aids in disease early detection. Cardiovascular disease, diabetes, liver disease, Alzheimer's disease, and Parkinson's disease are the most expensive diagnosis diseases.

It's a major challenge in the medical or healthcare industries to offer the highest quality services to all patients, and only those who can afford it can benefit from it. There is a vast amount of healthcare data available that is not being mined in a more efficient and reliable manner to uncover secret knowledge for successful decision-making. The proposed framework employs data mining techniques to detect Chronic diseases early. Machine learning is the process of programming computers to improve their output based on examples or previous data. The study of computer systems that learn from data and experience is known as machine learning. Training and Testing are the two stages of the machine learning algorithm. Prediction of a disease based on the signs and medical history of the patient Machine learning has been a stumbling block for decades. Machine Learning technology provides a strong forum in the medical sector for efficiently resolving healthcare issues.

II.RESEARCH OBJECTIVE

There is a need to research and develop a system that will enable end users to predict chronic diseases without having to visit a physician or doctor for diagnosis. To identify various diseases by observing the symptoms of patients and applying various Machine Learning Models techniques. There is no proper procedure for handling text and structured data. Both structured and unstructured data would be considered by the proposed framework. Machine Learning can improve the accuracy of predictions.

III.LITERATURE REVIEW

The study for the best medical diagnosis mining technique was performed by K.M. Al-Aidaroos, A.A. Bakar, and Z. Othman. For this study, the authors compared Nave Baeyes to five other classifiers: LR, KStar (K*), Decision Tree (DT), Neural Network (NN), and a basic rule-based algorithm (ZeroR). The efficiency of all algorithms was evaluated using 15 real-world medical problems from the UCI machine learning repository (Asuncion and Newman, 2007). In the experiment, NB outperformed the other algorithms in 8 of the 15 data sets, leading to the conclusion that the predictive accuracy results in Nave Baeyes are superior to other techniques. Darcy A. Davis, Nitesh V. Chawla, Nicholas Blumm, Nicholas Christakis, and Albert-Laszlo Barabasi discovered that treating chronic illness at a global level is neither time nor cost effective. As a result, the authors performed this study in order to forecast potential disease risk. CARE (which uses only a patient's medical history and ICD-9-CM codes to predict possible disease risks) was used for this. Based on their own medical history and that of similar patients, CARE incorporates collective filtering approaches with clustering to predict each patient's greatest disease risks. ICARE, an iterative version that integrates ensemble principles for improved efficiency, has also been defined by the authors. These cutting-edge systems don't need any advanced knowledge and can predict a wide range of medical conditions in a single run. ICARE's remarkable potential risk coverage means more precise early alerts for thousands of illnesses, several years ahead of time. When used to its full extent, the CARE system can be used to investigate a wider range of disease backgrounds, raise previously unconsidered questions, and facilitate discussions regarding early detection and prevention.

This research paper was written by JyotiSoni, Ujma Ansari, Dipesh Sharma, and SunitaSoni to provide a survey of existing techniques of information discovery in databases using data mining techniques that are used in today's medical research, specifically in Heart Disease Prediction. A number of experiments have been carried out to compare the performance of predictive data mining techniques on the same dataset, and the results show that Decision Tree outperforms, with Bayesian classification having comparable accuracy to Decision Tree in some cases, but other predictive approaches such as KNN, Neural Networks, and Classification based on Clustering underperform.

Shadab Adam Pattekari and Asma Parveen conducted a study to predict heart diseases using the Decision Tree Algorithm, in which the consumer provides data that is compared to a qualified set of values. As a result of this study, patients were able to provide basic information that was compared to data, and heart disease was expected. M.A.NisharaBanu and B. Gomathy analysed the various types of heart-related problems using medical data mining techniques such as association rule mining, grouping, and clustering I. The aim of a decision tree is to show any possible outcome of a decision. To achieve the best result, various rules are devised. The criteria used in this study were age, sex, smoking, being overweight, drinking alcohol, blood sugar, heart rate, and blood pressure. The risk level for various parameters is saved with their ids ranging from 1 to 100. (1-8). The standard level of prediction is represented by IDs less than 1, whereas higher IDs other than 1 represent higher risk levels. The pattern in the dataset is studied using the K-means clustering method. The algorithm divides the data into k groups. The closed cluster is allocated to each point in the dataset. Each cluster centre is recalculated as the average of the cluster's points.

IV PROPOSED SYSTEM

We have mixed structured and unstructured data in the healthcare fields to determine disease risk in this project. The use of a latent factor model to recreate missing data in medical records obtained from online sources. We could also assess the major chronic diseases in a specific area and population using statistical information. We consult hospital experts to learn about useful features when dealing with structured data. In the case of unstructured text files, we use the randrom forest algorithm to automatically select features.

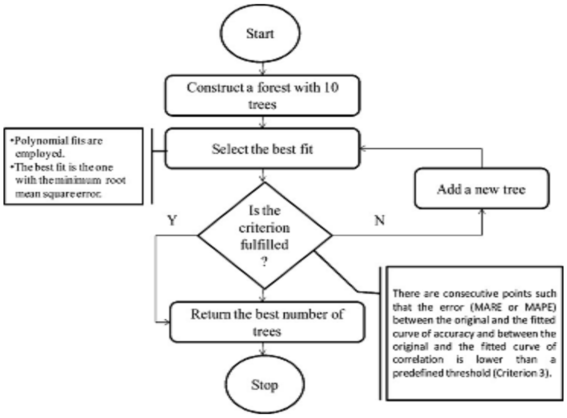

Start
Construct a forest with 10 trees
Select the best fit
•Polynomial fits are employed.
•The best fit is the one with the minimum root mean square error.
Add a new tree
Is the criterion fulfilled ?
Y
N
There are consecutive points such that the error (MARE or MAPE) between the original and the fitted curve of accuracy and between the original and the fitted curve of correlation is lower than a predefined threshold (Criterion 3).
Return the best number of trees
Stop

Fig 1: - System Model

4.1 Data collection

Data collection has been done from the internet to identify the disease here the real symptoms of the disease are collected i.e. no dummy values are entered. The symptoms of the disease are collected from different health related websites.

4.2 Data Preprocessing

Before feeding the data into the Prediction model, following data cleaning and preprocessing steps are performed

- Checking null values and filling using forward fill method
- Converting data into different cases
- Standardizing the data using mean and standard deviation
- Splitting the dataset into training and testing sets

4.3 Building Model

Many methods are used to perform data mining. Machine learning is one of the approaches. Random forest Machine learning strategies include grouping, clustering, summarization, and many others. Since classification techniques are used in this project, classification is one of the data mining processes in this phase of categorical data classification. And this step is divided into two phases: training and testing. In the training phase, predetermined data and associated class labels are used for classification. The training stage is often referred to as supervised learning. The preparation and testing phases of the classification process are depicted in the diagram. In the training process, training tuples are used, and in the test data phase, test data tuples are used, and the classification rule's accuracy is calculated. Assume that the classification rule's accuracy on testing data is sufficient for the rule to be used for classification of unmined data.

4.4 Prediction:

Prediction using Random Forest: -

Prediction done by Random Forest Model using Flask framework model trained by training chronic disease dataset

4.5Algorithm

4.4.1 Random Forest Algorithm

Input: Dataset
Output: Predicted class label
Step 1 : Set Number of classes = N, Number of features =M
Step 2 : Let „m‟ determine the number of features at a node of decision tree, (m < M)
Step 3 : For each decision tree do
Select randomly: a subset (with replacement) of training data that represents the N classesand use the rest of data to measure the error of the tree
Step 4 : For Each node of this tree do Select randomly: m features to determine the decision at this node and calculate the best split accordingly.
Step 5: End for
Step6 : End For

4.2.2 CNN(CONVOLUTIONAL NEURAL NETWORK):

A convolutional neural network is a feed-forward neural network that is generally used to analyze visual images by processing data with grid-like topology. It’s also known as a ConvNet. A convolutional neural network is used to detect and classify objects in an image. A convolution neural network has multiple hidden layers that help in extracting information from an image. The four important layers in CNN are:

1. Convolution layer
2. ReLU layer
3. Pooling layer
4. Fully connected layer Convolution Layer

This is the first step in the process of extracting valuable features from an image. A convolution layer has several filters that perform the convolution operation. Every image is considered as a matrix of pixel values

4.2.2.1ReLU layer

ReLU stands for the rectified linear unit. Once the feature maps are extracted, the next step is tomove them to a ReLU layer. ReLU performs an element-wise operation and sets all the negative pixels to 0. It introduces non-linearity to the network, and the generated output is a rectified feature map.

4.2.2.2 Pooling Layer

Pooling is a down-sampling operation that reduces the dimensionality of the feature map. The rectified feature map now goes through a pooling layer to generate a pooled feature map.

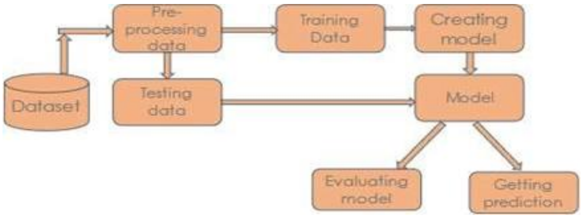


FIG 2: SYSTEM ARCHITECTURE

V.RESULTS AND CONCLUSION

Model	Accuracy
Diabetes Model	98.25%
Breast Cancer Model	98.25%
Heart Disease Model	85.25%
Kidney Disease Model	99%
Liver Disease Model	78%
Malaria model (CNN)	96%
Pneumonia model (CNN)	95%

Table 1: - shows the accuracy achieved using random forest for each disease

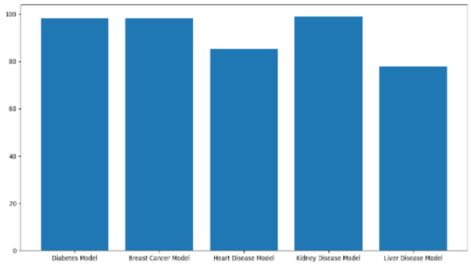
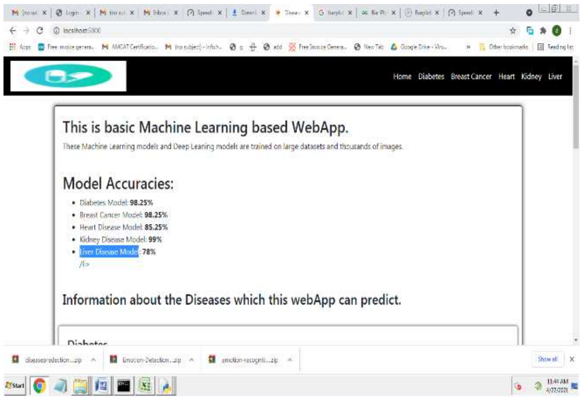


Fig. 3 shows the accuracy or each model using Random forest classifier



This is basic Machine Learning based WebApp.

These Machine Learning models and Deep Leaning models are trained on large datasets and thousands of images.

Model Accuracies:

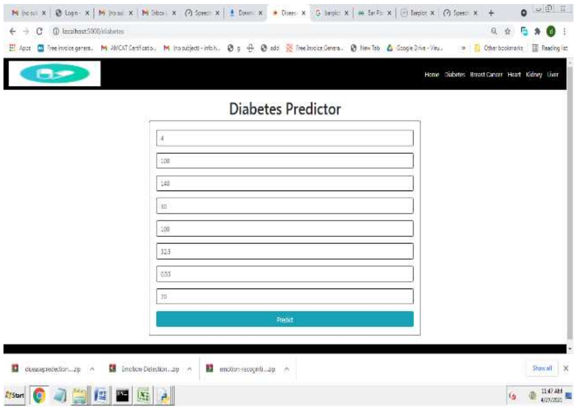
- Diabetes Model: 98.25%
- Breast Cancer Model: 98.25%
- Heart Disease Model: 85.25%
- Kidney Disease Model: 99%
- Liver Disease Model: 78%

/li>

Information about the Diseases which this webApp can predict.

Diabetes

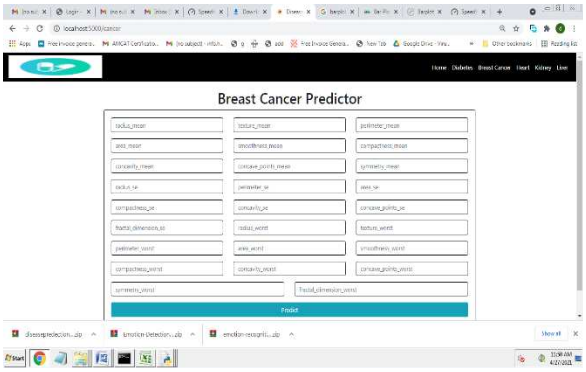
Fig. 4 Home screen of the prediction system



Diabetes Predictor

4
100
140
80
100
32.5
0.55
30

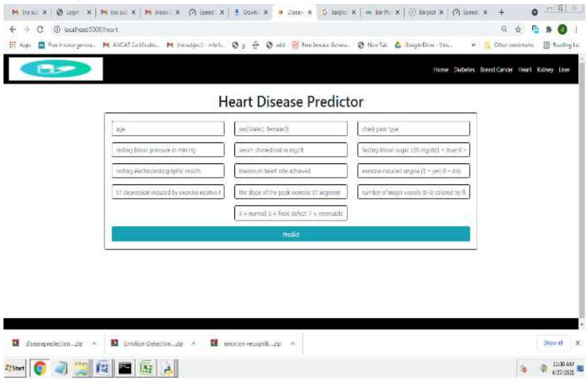
Fig. 5 :- Diabetes Prediction entry form



Breast Cancer Predictor

radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave_points_mean
symmetry_mean
radius_se
perimeter_se
area_se
compactness_se
concavity_se
concave_points_se
fractal_dimension_se
radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave_points_worst
symmetry_worst
fractal_dimension_worst

Fig. 6 :- Breast cancer Prediction entry form



Heart Disease Predictor

age
sex(Male:1, female:0)
chest pain type
resting blood pressure in mm Hg
serum cholestoral in mg/dl
fasting blood sugar 120 mg/dl(1 = true; 0 =
resting electrocardiographic results
maximum heart rate achieved
exercise induced angina (1 = yes; 0 = no)
ST depression induced by exercise relative t
the slope of the peak exercise ST segment
number of major vessels (0-3) colored by fl
3 = normal; 6 = fixed defect; 7 = reversable

Fig. 7 :- Heart Disease Prediction entry form

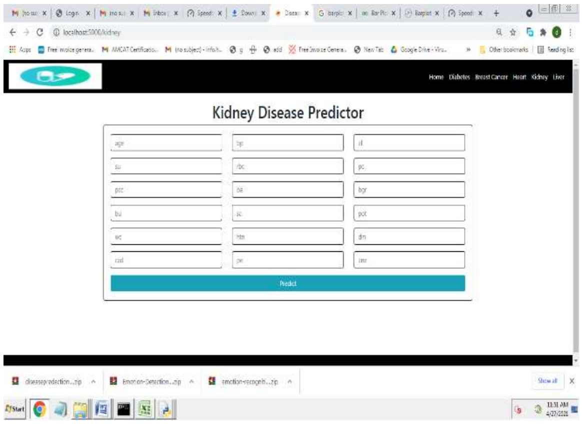


Fig. 8:- Kidney Disease Prediction entry form

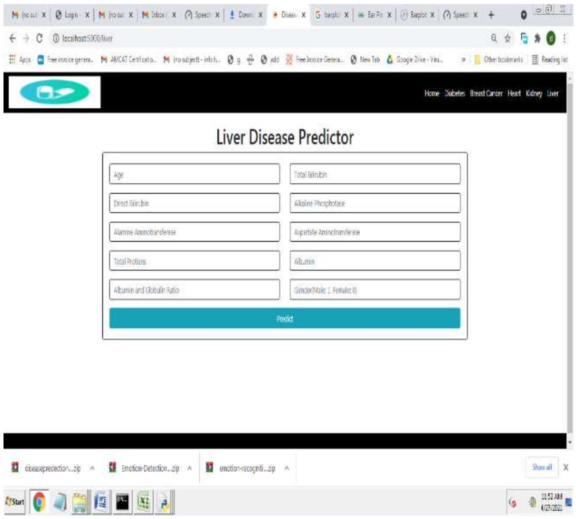
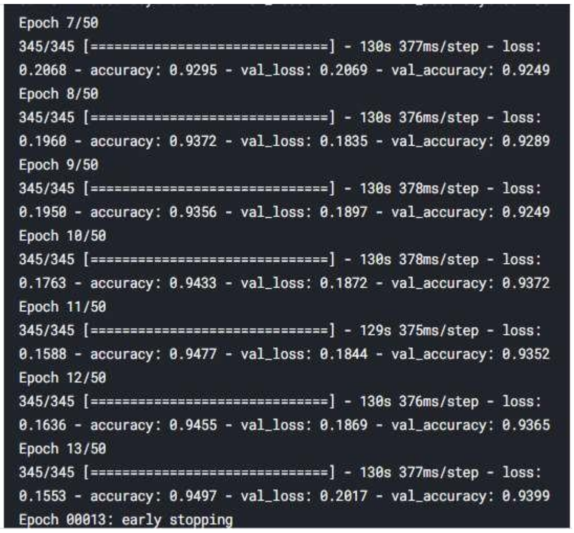


Fig. 9:- Liver Disease Prediction entry form



```
Epoch 7/50
345/345 [==============================] - 130s 377ms/step - loss:
0.2068 - accuracy: 0.9295 - val_loss: 0.2069 - val_accuracy: 0.9249
Epoch 8/50
345/345 [==============================] - 130s 376ms/step - loss:
0.1960 - accuracy: 0.9372 - val_loss: 0.1835 - val_accuracy: 0.9289
Epoch 9/50
345/345 [==============================] - 130s 378ms/step - loss:
0.1950 - accuracy: 0.9356 - val_loss: 0.1897 - val_accuracy: 0.9249
Epoch 10/50
345/345 [==============================] - 130s 378ms/step - loss:
0.1763 - accuracy: 0.9433 - val_loss: 0.1872 - val_accuracy: 0.9372
Epoch 11/50
345/345 [==============================] - 129s 375ms/step - loss:
0.1588 - accuracy: 0.9477 - val_loss: 0.1844 - val_accuracy: 0.9352
Epoch 12/50
345/345 [==============================] - 130s 376ms/step - loss:
0.1636 - accuracy: 0.9455 - val_loss: 0.1869 - val_accuracy: 0.9365
Epoch 13/50
345/345 [==============================] - 130s 377ms/step - loss:
0.1553 - accuracy: 0.9497 - val_loss: 0.2017 - val_accuracy: 0.9399
Epoch 00013: early stopping
```

Fig 10 :- this graph represents the training and validation accuracy of malaria disease where the training accuracy is 94.7 at 13 epochs and Val accuracy is 93.99

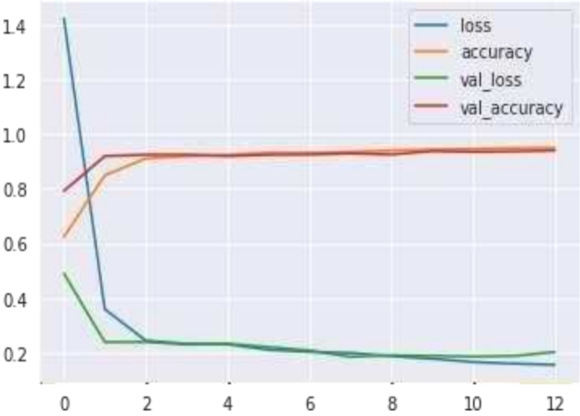
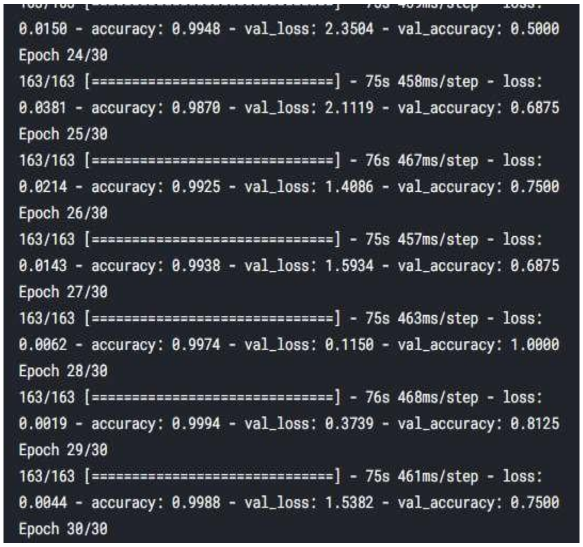


Fig 11 :-this graph represents the Malaria disease training and validation accuracy and training and validation loss



```
0.0150 - accuracy: 0.9948 - val_loss: 2.3504 - val_accuracy: 0.5000
Epoch 24/30
163/163 [==============================] - 75s 458ms/step - loss:
0.0381 - accuracy: 0.9870 - val_loss: 2.1119 - val_accuracy: 0.6875
Epoch 25/30
163/163 [==============================] - 76s 467ms/step - loss:
0.0214 - accuracy: 0.9925 - val_loss: 1.4086 - val_accuracy: 0.7500
Epoch 26/30
163/163 [==============================] - 75s 457ms/step - loss:
0.0143 - accuracy: 0.9938 - val_loss: 1.5934 - val_accuracy: 0.6875
Epoch 27/30
163/163 [==============================] - 75s 463ms/step - loss:
0.0062 - accuracy: 0.9974 - val_loss: 0.1150 - val_accuracy: 1.0000
Epoch 28/30
163/163 [==============================] - 76s 468ms/step - loss:
0.0019 - accuracy: 0.9994 - val_loss: 0.3739 - val_accuracy: 0.8125
Epoch 29/30
163/163 [==============================] - 75s 461ms/step - loss:
0.0044 - accuracy: 0.9988 - val_loss: 1.5382 - val_accuracy: 0.7500
Epoch 30/30
```

Fig 12: - Pneumonia disease accuracy here we can see that training accuracy is 99.88 and validation accuracy is 75.00 after 30epochs and training loss is 0.0044 and validation loss is 1.5382

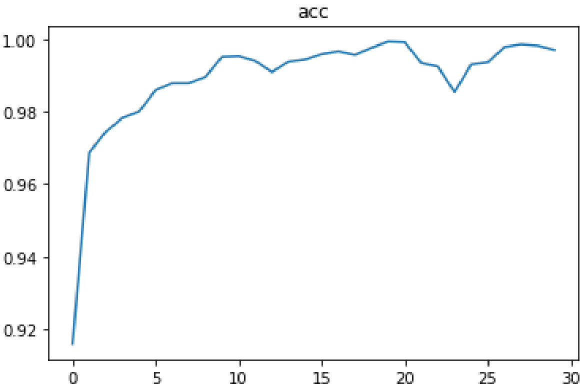


Fig 13: - this figure represents the Pneumonia Training accuracy graph the accuracy of the model is 99.88

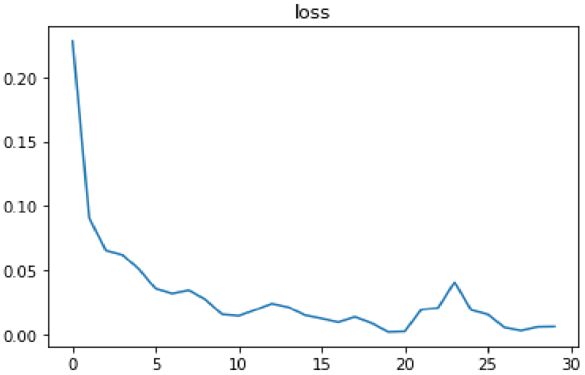
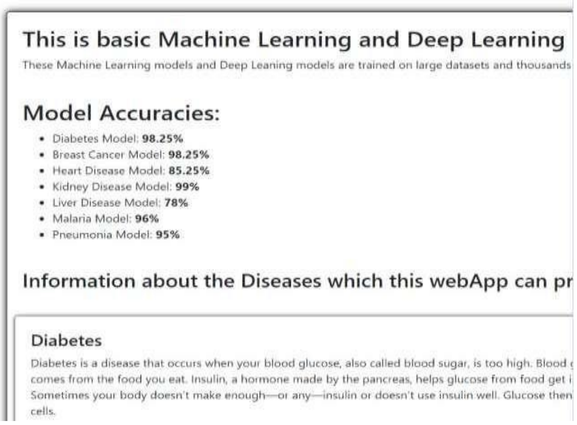


Fig 14: - this figure represents the Pneumonia Training loss graph and the loss of the training data is 0.0044



This is basic Machine Learning and Deep Learning

These Machine Learning models and Deep Leaning models are trained on large datasets and thousands

Model Accuracies:

- Diabetes Model: **98.25%**
- Breast Cancer Model: **98.25%**
- Heart Disease Model: **85.25%**
- Kidney Disease Model: **99%**
- Liver Disease Model: **78%**
- Malaria Model: **96%**
- Pneumonia Model: **95%**

Information about the Diseases which this webApp can pr

Diabetes

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood comes from the food you eat. Insulin, a hormone made by the pancreas, helps glucose from food get Sometimes your body doesn't make enough—or any—insulin or doesn't use insulin well. Glucose then cells.

Fig 12 this figure represents home page of the web site Which gives the different discerption of the diseases and accuracy obtained from each disease

VII. CONCLUSION

The aim of this project is to predict disease based on symptoms. The project is set up in such a way that the device takes the user's symptoms as input and generates an output, which is disease prediction. A prediction accuracy probability of 95% is obtained on average. The grails system was used to successfully incorporate Disease Predictor.

REFERENCE

[1] A.Davis, D., V.Chawla, N., Blumm, N., Christakis, N., & Barbasi, A. L. (2008). Predicting Individual Disease Risk Based on Medical History.

[2] Adam, S., & Parveen, A. (2012). Prediction System for Heart Disease Using Naive Bayes.

[3] Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical Data Classification with Naive Bayes Approach. *Information Technology Journal* .

[4] Darcy A. Davis, N. V.-L. (2008). Predicting Individual Disease Risk Based on Medical History.

[5] JyotiSoni, Ansari, U., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction.

[6] K.M. Al-Aidaroos, A. B. (n.d.). K.M. Al-Aidaroos, A. B. (n.d.). 2012. *Medical Data Classification with Naive Bayes Approach* .

[7] NisharBanu, MA; Gomathy, B.; (2013). Disease Predicting System Using Data Mining Techniques.