International Journal of Innovative Research in Engineering
Volume 4, Issue 2 (March-April 2023), PP: 119-126.
www.theijire.com ***ISSN No: 2582-8746***

Identifying Multiple DiseaseUsing Machine Learning

Supriya U[1], Mukunth Balaji J[2], Deepak Kumar J[3]
[1] Assistant Professor, Dept. of Computer Science and Business System, Bannari Amman Institute of Technology, TN, India.
[2,3]Dept. of Computer Science and Business System, Bannari Amman Institute of Technology, TN, India.

How to cite this paper:
Supriya U[1], Mukunth Balaji J[2], Deepak Kumar J[3]. "Identifying Multiple DiseaseUsing Machine Learning", IJIRE-V4I02-119-126.

Abstract: *Early disease identification may assist to reduce the number of fatalities. In recent years, some researchers have applied various machine learning-based methodologies to construct autonomous disease detection systems. The goal of the disease detection models is to integrate the domains of medicine and artificial intelligence (AI) so that people can see how effectively they can complement one another. We intend to carry out an extensive study on AI applications for the healthcare industry to better grasp the function of AI in the medical profession. We'll start by reviewing the highlights and justifications for employing AI in the healthcare sector. Then, we thoroughly discuss machine-learning-based integration methods for the healthcare industry. After discussing the technical issues with AI in the medical sector, we then demonstrate how machine learning might be useful. We also investigate how machine learning is affecting the medical industry. Also, we offer a number of noteworthy initiatives that highlight the value of machine learning in healthcare applications and services. Lastly, talk about some current problems with diagnosing diseases and recommend future research and development directions that will result in the application of machine learning in the healthcare industry.*

Key Word: *Random Forest, SVM,CNN, Machine Learning*

I.INTRODUCTION

Recent changes in people's lifestyles and socioeconomic circumstances have been influenced by swift technological advancements and the healthcare sector, increasing the likelihood that people will catch a variety of diseases. major illnesses including lung cancer and brain tumours. Among other things, pneumonia and cancer affect the entire world. According to the World Health Organization, around 86000 persons had brain tumours in 2019 with an average survival rate of 35% , while lung cancer is a horrible illness that kills one in every five people worldwide, or 1.59 million people, accounting for 19.4% of all fatalities. The coronavirus pandemic, which has affected many nations and caused more than 37 million confirmed cases and more than 1 million fatalities worldwide, has raised awareness of diseases like pneumonia. These serious illnesses raise society demands and healthcare costs, which have an effect on the patient's general health. Determining a person's risk of getting one or more serious diseases is the main objective of disease detection. This calls for the careful analysis of a wide range of concerns, which consumes a substantial amount of time and money.Many medical institutes across the world can now easily acquire medical datasets for health-related data. Examples of several types of medical data include image data, patient reports, and others. Medical data is particularly difficult to manage since it is very erratic, irregular, and contains unstructured data. Based on avariety of factors, including the patient's medical condition, the level of the doctor, and the variations in patient reports, among others, manual data entry is impossible and diagnosisis constrained.

By including a machine learning-based disease detection module to help with disease prediction and diagnosis, these difficulties are solved.. Also, it can be challenging for the user if they are far from medical facilities because the condition cannot be identified. In light of the foregoing Using automated software to do the treatment can save time and money, be better for the patient, and make the process go more smoothly. Other Heart Disease Prediction Systems analyse the patient's risk level using data mining techniques. A web-based tool called Disease Predictor can identify a user's ailment based on the symptoms they exhibit. The Illness Prediction system has collected data sets from many health-related websites.

With Disease Predictor, the customer will be able to assess the possibility of a condition based on the provided symptoms.Individuals are constantly interested in learning new things, especially given how popular the internet is becoming. When a problem arises, people frequently desire to search the internet for it. Compared to the broader public, hospitals and doctors have less access to the internet. When someone has a disease, they are limited in their options. So, this system may be advantageous to people. Long-lasting or slow-healing disorders are referred to as chronic illnesses, and many of these conditions can only be managed with regular therapies rather than being healed. Similar to other countries, India is going through major social and economic change, which is hastening the growth in cardiovascular disease prevalence.

All countries currently face a problem with chronic diseases, which affect one-third of the population in each. Care for chronic diseases is more expensive, and it is challenging for the sick. A lot of chronic disease datasets are gathered and analysed in the medical area, and data mining helps with disease early detection. The most expensive diagnoses include heart disease, diabetes, breast cancer, and Parkinson's disease. Offering the finest quality services to all patients in the medical or healthcare fields is a significanttask, and only those who can afford it can profit from it. There is a tonne of healthcare data available, but it isn't being mined in a way that's more effective and dependable. Discover hidden information to make wise

decisions. Data mining techniques are used in the proposed framework to find chronic diseases early. Programming computers to produce better results based on examples or past data is known as machine learning. Machine learning is the study of computer systems that learn from information and experience. The machine learning algorithm has two stages: training and testing. illness prognosis based on patient symptoms and medical history For many years, machine learning has been a challenge. In the medical industry, machine learning technology offers a powerful platform for quickly resolving healthcare-related problems. Similar to human brains, machine learning models are given data and extract distinguishing features, simulating cognitive functions like vision. It mimics medical professionals when diagnosing diseases and builds experience over time throughconstant practice to improve detection accuracy and hence strengthen the model's resilience. The application has produced outstanding outcomes.

II.RESEARCH OBJECTIVE

A system that will enable end users to forecast chronic diseases without needing to see a doctor or medical professional for a diagnosis needs to be researched and developed. by observing patient symptoms and using a variety of machine learning modeling techniques, different diseases can be identified. Text and structured data handlingdo not follow any standard procedure.

The suggested framework would consider both organized and unstructured data. Prediction accuracy can be increased by machine learning.

III.LITERATURE REVIEW

1. According to the paper focuses about as diabetes is one of the dangerous diseases in the world , it can cause many varieties of disorders which includes blindness etc .In this paper they have used machine learning techniques to find out diabetes disease as it is easy and flexible to forecastwhether the patient has illness or not . Their aim of this analysis was to invent a system that can help the patient to detect the diabetes disease of the patient with accurate results. Decision Tree, Nave Bayes, and SVM algorithms were the major four utilised here, and they compared their accuracy, which was 85%, 77%, and 77.3%, respectively. They also used ANN algorithm after the training process to see the reactions of the network which states whether the disease is classified properly or not . Herethey compared the precision recall and F1 score support ad accuracy of all the models[1] .

2. The main aim of the paper is ,as heart plays an important role in living organisms. So the diagnosis and prediction of heart related disease should be perfect and correct because it is very crucial which can cause death cases related to heart .Artificial intelligence and machine learning thus aid in the prediction of all kinds of natural disasters. So, they use the UCI repository dataset for training and testing in this research to calculate the accuracyof machine learning for predicting heart disease using k-nearest neighbour, decision tree, linear regression, and SVM. They also compared SVM 83%, Decision Tree 79%, Linear Regression 78%, and K-Nearest Neighbor 87% in terms of algorithm accuracy[2].

3. According to the system, liver illnesses are a major cause of death in India and are regarded as a serious illness worldwide. thus early liver disease detection is challenging. Hence, we can accurately diagnose liver illness using an automated software that uses machine learning methods..They used and compared SVM ,Decision Tree andRandom forest algorithm and measures precision, accuracy and recall metrics for quantitative measurement. Theaccuracy are 95%,87%,92% respectively[3].

4. According to [4], The current surveys, and also latest machine learning-based approaches for tumorcategorization, were thoroughly examined. The survey covers the basic methods of machine learning-based brain tumor categorization techniques such Data preprocessing, extraction of features, and categorization, as well as their accomplishments and limitations.

5. According to [5], include the addition of invasion as a list of requirements for atypical meningioma, as well as the inclusion of a soft tissue grading system for theChanges include the newly combined entity of isolated fibrous tumour hemangiopericytoma, which deviates from how other CNS tumours are rated. In general, the 2016 CNS WHO aims to support medical, scientific, and epidemiologic research that will prolong the lives of persons with brain tumours. The majority of partners had widespread dysfunction for months before the symptom that prompted adoctor's visit.

6. The suggested method is extremely accurate and effective at diagnosing, classifying, and segmenting brain tumours, claims [6]. This calls for the application of precise automatic or semi-automatic approaches. The paper offers a CNN-based automatic segmentation technique that locates small 3x3 kernels (Convolution Neural Networks). Combining these two methods can be used to segment and classify data. The machine learning technology NN, which uses layers to predict outcomes, gave rise to CNN (Neural Networks). The proposed approaches comprise phases for data collection, data pre-processing, filtering, segmentation, extraction of features, classification using Convolutional Neural Networks, and identification. It is possible to extract significant patterns and relationships from data using data mining techniques.

7. In line with [7], The dataset for Pima Indians was subjected to the algorithm. The authors made no use of any pre-processing methods. 90% of the dataset is used as the training set, while 10% is used as the testing set. The proposed technique attained accuracies of 89.56%, 81.49% for the training and testing data, respectively.

IV.PROPOSED SYSTEM

For this project, we combined structured and unstructured data from the healthcare sectors to assess illness risk. the creation of missing data using a latent component model to be applied to online-sourced medical records. Statistical data could also be used to evaluate the main chronic diseases in a particular region and population.To find out about practical aspects for working with structured data, we speak with hospital specialists. We employ the randrom forest method to automatically choosefeatures for unstructured text files.

4.1 Data collection

To identify the disease, data was gathered from the internet. No dummy values were entered; instead, the true symptomsof the disease were collected. The disease's symptoms are gathered from a variety of health-related sources..

4.2 Data Preprocessing

The following data cleaning and preparation activities are carried out prior to incorporating the data into the prediction model.

- Examining null values and using the forward fillmethod to fill.
- transforming data into several scenarios.
- Using the mean and standard deviation, standardize the data.
- creating training and test sets from the dataset

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.

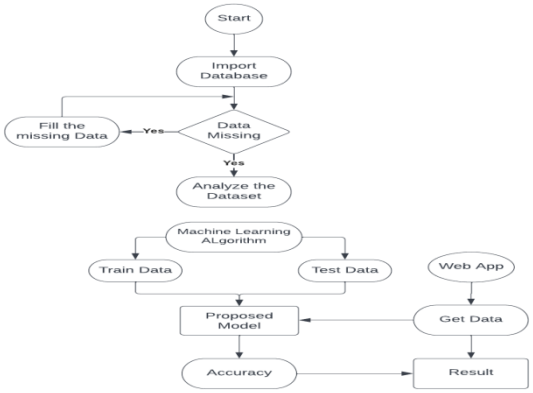


Figure:1 - Proposed Model

The proposed model of the multiple disease prediction is inFigure:1.

4.4Algorithms

4.4.1.K-Nearest Neighbor Algorithm

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictionsabout the grouping of an individual data point.

The working of the K-NN algorithm is as followed:

- Step 1: Decide on the neighbours' K-numbers.
- Compute the Euclidean distance between Kneighbours in step two.
- Step 3: Based on the determined Euclidean distance,select the K closest neighbours.
- Step 4: Count the number of data points in eachcategory among these k neighbours.
- Step 5: Allocate the fresh data points to the categorywhere the neighbour count is highest.
- Step 6: Our model is complete.

4.4.2 Random Forest Algorithm

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of thatdataset.

The working of the Random Forest algorithm is as followed:

- Step 1: Choose K data points at random from thetraining set.
- Step 2: Construct the decision trees linked to thechosen data points (Subsets).
- Step 3: Choose N for the size of the decision treesyou wish to construct.

- Repeat steps 1 and 2 in step 4.
- Step 5: Assign new data points to the category thatreceives the majority of votes by looking up each decision tree's predictions for the new data points.

4.4.3. Support Vector Machine Algorithm

In order to make it simple to place fresh data points in the appropriate category in the future, the Support Vector Machine technique is used to generate the best line or decision boundary that can divide n-dimensional space into classes. A hyper plane is the name given to this optimal decision boundary. The extreme points and vectors that aid in the creation of the hyper plane are selected by the support vector machine. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method.

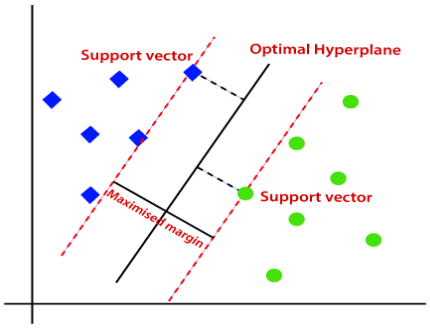


Figure:2 - SVM

In Figure:2, Although H2 has the smallest marginal width, it classifies the data points while H1 does not. Given that it correctly classifies the data points and has the largest marginal width, the hyper plane H3 is the best or ideal classifier. SVM is mostly used to linearly separate the classes of the output variable, to put it simply. The SVM method assists in identifying the appropriate decision boundary or region, often known as a hyper plane. The SVM algorithm determines which line from each class is closest to the other. Support vectors are the names for these points.

4.4.4Linear Regression

The linear regression algorithm, often known as linear regression, demonstrates a linear relationship between a dependent (y) and one or more independent (y) variables. Given that linear regression demonstrates a linear relationship, it may be used to determine how the dependent variable's value changes as a function of the independent.

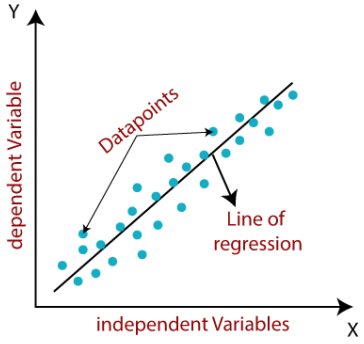


Figure: 3- Linear Regression

In Figure: 3, y= a0+a1x+ ε
Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error
The values for x and y variables are training datasets for Linear Regression model representation.

4.4.5. Convolutional Neural Network (CNN)

A feed-forward neural network called a convolutional network analyses visual images by processing data in a grid-like architecture. It is also referred to as a ConvNet. To find and categorise items in an image, a convolutional neural network is employed. Many hidden layers in a convolution neural network aid in information extraction. The four important layers in CNN:

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

4.4.5.1. ReLU Layer:

ReLU stands for the rectified linear unit. The next step is to transfer the feature maps to a ReLU layer after they have

been retrieved. ReLU executes an elementwise operation, setting all of the dark pixels to 0.

The result is a corrected feature map, and it gives thenetwork non-linearity.

4.4.5.2Pooling Layer: Pooling is a downsampling techniquethat lowers the feature map's dimensionality. To create a pooled feature map, the rectified feature map is now passed through a layer of pooling.

The Convolutional Neural Network explains how ReLU layer and Pooling layer function together is illustrated belowin Figure: 4.

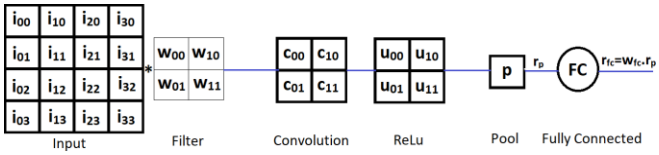


Figure: 4

V.PROPOSED ARCHITECTURE

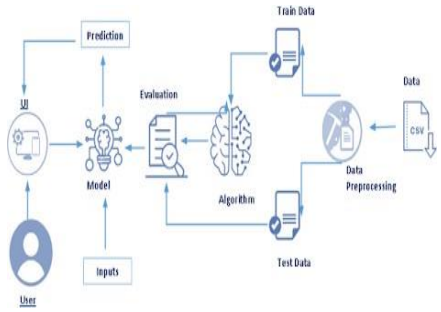


Figure: 5 System Architecture and working flowof the model

VI.RESULT

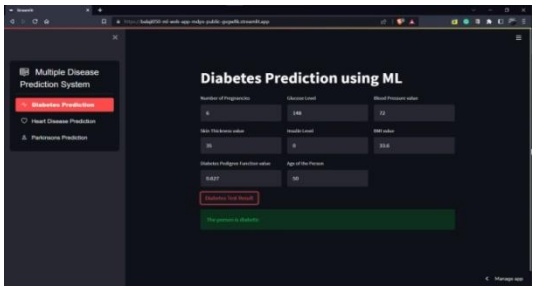


Figure 6: this figure represents Creating of DiabetesPrediction webpage using streamlit.

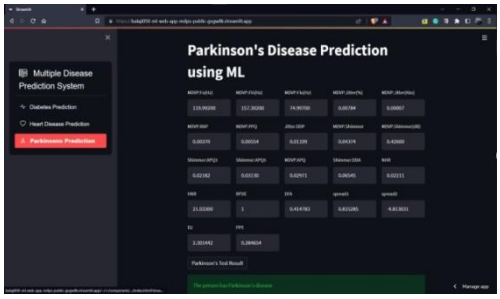


Figure 7: this figure represents Streamlit Web Application for Parkinson disease and values are been entered and predicted as The person is having Parkinson's disease

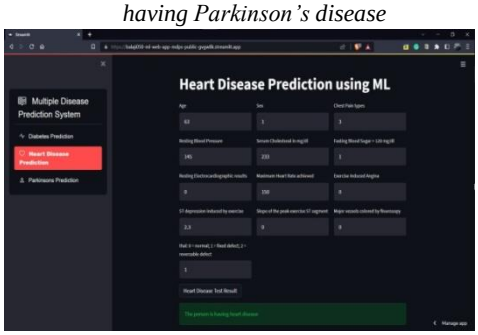


Figure 8: this figure represents Heart Disease prediction web page created in a streamlit application which is hosted in streamlit cloud and this web page have predicted as The person is having Heart Disease with the entered value

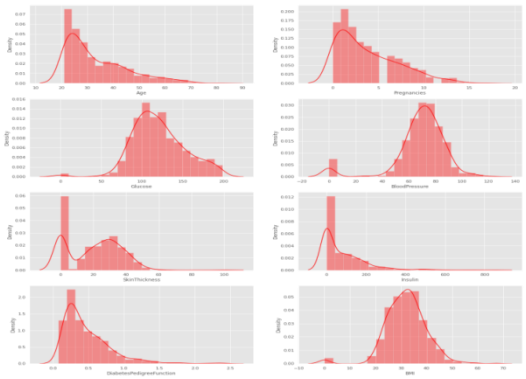


Figure 9: this figure represents Ploting of graph on different parameters that are present in the dataset and the value present or the weightage of the present dataset

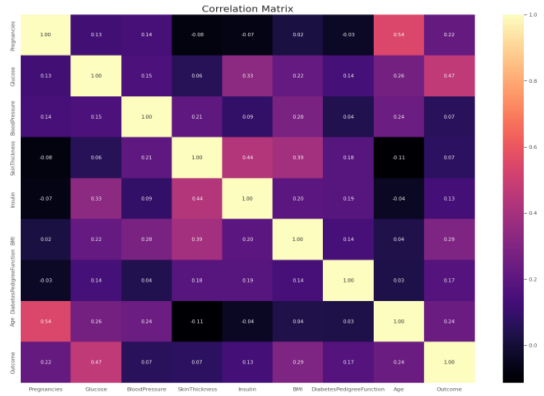
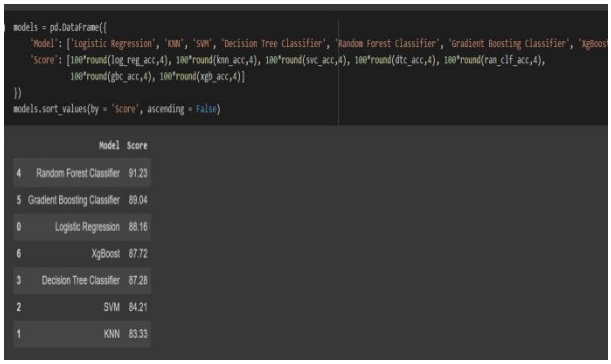


Figure 10: this figure represents Confusion matrix on Diabetes dataset by using this confusion matrix we can understand that in some case dataset is well balanced for prediction.



```
models = pd.DataFrame({
    'Model': ['Logistic Regression', 'KNN', 'SVM', 'Decision Tree Classifier', 'Random Forest Classifier', 'Gradient Boosting Classifier', 'XgBoost
    'Score': [100*round(log_reg_acc,4), 100*round(knn_acc,4), 100*round(svc_acc,4), 100*round(dtc_acc,4), 100*round(ran_clf_acc,4),
              100*round(gbc_acc,4), 100*round(xgb_acc,4)]
})
models.sort_values(by = 'Score', ascending = False)
```

	Model	Score
4	Random Forest Classifier	91.23
5	Gradient Boosting Classifier	89.04
0	Logistic Regression	88.16
6	XgBoost	87.72
3	Decision Tree Classifier	87.28
2	SVM	84.21
1	KNN	83.33

Figure 11: this figure represents Module Comparison on different algorithms used on Parkinson's disease model and we can conclude that by using random forest classifier we can get better accuracy for prediction on that dataset

Figure 12: this figure represents plotting graphs on different parametes in heart disease prediction dataset which dataset Weight are represented in an dotted graphs.

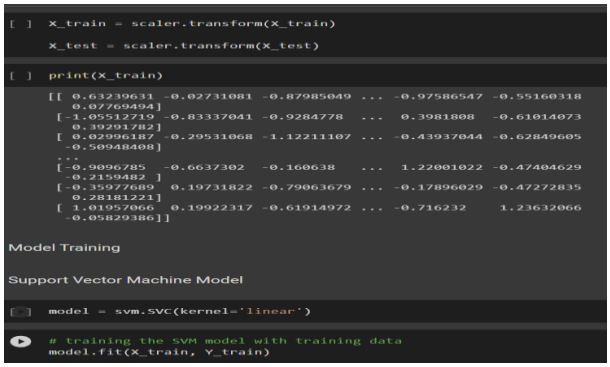


Figure 13: this figure represents data Scaling which done to get an better accuracy and by using this techniques we can reduce the loss value to get better dataset for the training result in the model.

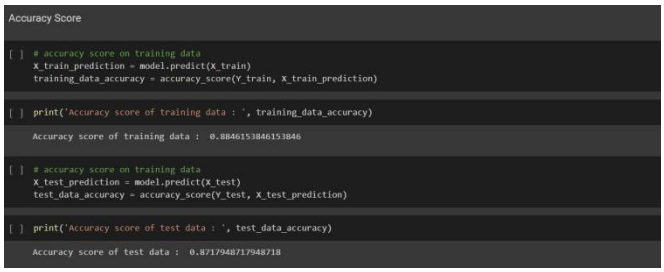


Figure 14: this figure represents Accuracy Score Generated for Parkinson's models with an training data score of 88.46 and testing data score of 87.17.

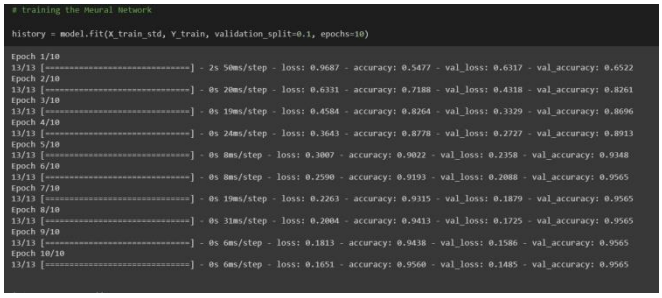


Figure 15: This graph represents the training and validation accuracy of Breast Cancer disease where the training accuracy is 95.60 at 10 epochs and Val loss is 14.8

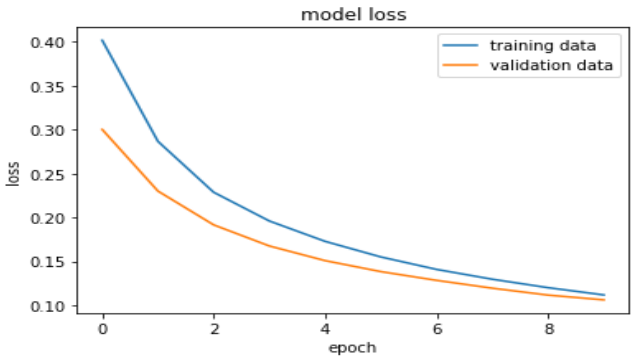


Figure 16: this figure represents the Brest Cancer Training loss graph and the loss of the training data and validationdata is less than 12

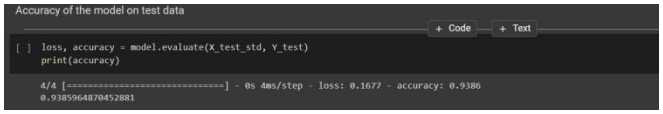


Figure 17: this figure represents Accuracy level of themodel breast cancer with an accuracy of 93.85

VII.CONCLUSION

The future of medical healthcare has more contemporaryprospects against the background of artificial intelligence and machine learning techniques. Due to its distinct feature processing method and changeable model structure, machine learning has become a key factor in future advancement in the face of unstable medical data. A more complicated machine learning system network is created as a result of the machine learning models being linked together and learning from one another. This system network aids in the development of medical diagnostics and practical applications, which advances the medical

profession. In this analysis, we've highlighted the most popular machine learning techniques. The method, current difficulties, and machine learning's restrictions are also presented. We also discuss the numerous security and privacy concerns as well as the difficulties that have been faced. We also highlight a number of pertinent studies and other research issues that need to be further addressed.

References

1. *Priyanka Sonar, Prof. K. JayaMalini," DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES", 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication (ICCMC).*
2. *Archana Singh ,Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3).*
3. *A.Sivasangari, Baddigam Jaya Krishna Reddy,Annamareddy Kiran, P.Ajitha," Diagnosis of Liver Disease using Machine Learning Models" 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).*
4. *Khan Muhammad; Salman Khan; Javier Del Ser; Victor Hugo C. de Albuquerque. "machine Learning for Multigrade Brain Tumor Classification in Smart Healthcare Systems: A Prospective Survey" IEEE Transactions on Neural Networks and Learning Systems (Volume: 32, Issue: 2, Feb. 2021).*
5. *Khan Muhammad; Salman Khan; Javier Del Ser; Victor Hugo C. de Albuquerque. "machine Learning for Multigrade Brain Tumor Classification in Smart Healthcare Systems: A Prospective Survey" IEEE Transactions on Neural Networks and Learning Systems (Volume: 32, Issue: 2, Feb. 2021).*
6. *G. Hemanth, M. Janardhan and L. Sujihelen, "Design and Implementing Brain Tumor Detection Using Machine Learning Approach", 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019.*
7. *Cao, C., Liu, F., Tan, H., Song, D., Shu, W. et al. (2018). Deep learning and its applications in biomedicine.Genomics, Proteomics & Bioinformatics, 16(1), 17–32. DOI 10.1016/j.gpb.2017.07.003*
8. *Dr. Ankita Karale, Uday Talpade, Sahil Nikumbh, Laxman Wadekar , Parikshit Angre, "Multiple Disease Detection Using Machine Learning: A Survey ", International Journal for Research in Applied Science & Engineering Technology,June 2022,ISSN: 2321-9653*
9. *Ankush Singh, Ashish Yadav, Saloni Shah ,Prof. Renuka Nagpure, "Multiple Disease Prediction System",International Research Journal of Engineering and Technology, Mar 2022 , ISSN: 2395-0056*
10. *Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016,*
11. *December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.*
12. *Aishwarya Mujumdar, Dr. Vaidehi V, "Diabetes Prediction using Machine Learning Algorithms",INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019,*
13. *ICRTAC 2019*
14. *Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.*