EasyChair Preprint

№ 7367

AI-DOC: a Mini Healthcare Assistant for the Digital World

Abhishek Parashar, Yukti Mohan and Sayoni Ghosh

January 25, 2022

AI-DOC: A mini healthcare assistant for the digital world

Abhishek Parashar [a]
ECE Department
parasharabhishek62@gmail.com

Yukti Mohan[a,b]
ECE Department
yukti.mohan99@gmail.com

Sayoni Ghosh[a]
ECE Department
sayonighosh2011@gmail.com

***Abstract*—In today's times, healthcare cost is increasing with each passing day but there are many instances when money and energy can be saved using the power of Artificial Intelligence. Data is available in huge amounts and can be used for our benefit to solve several problems. It is also seen that analyzing and accessing reports is cumbersome for most people. Moreover, many people lack the basic knowledge of several prevalent diseases. In the given work, the authors propose a solution to all the aforementioned problems, called the AI-DOC. It is an application that uses data and machine learning models and aims at providing a low-cost service of predicting diseases like breast cancer, diabetes, liver disease, malaria, pneumonia, and sepsis by anyone, anywhere. AI-DOC is created to make medical treatment hassle-free and available for all. COVID-19 pandemic has proved the importance of such applications as people dread to go out and consult a doctor. This application has the potential to help millions of people to self-diagnose and take necessary measures in time.**

***Index Terms*—KNN, CNN, Gradient Boosting, Linear Regression, Self Analysis**

I. Introduction

The year 2020 has forced us to realize the need and importance of healthcare applications to beget a more satisfying quality of life and improved health assistance. Until now, it was more convenient to visit a doctor than analyze the health reports yourself because a majority of the population is not well-versed with the technicalities and does not have the expertise of a doctor to interpret the numbers shown in a medical report [30]. Because of this, both web and mobile health applications have hardly made any success. On top of that, it has become compulsory to book an appointment with the doctor, or simply said, a meeting with the doctor is bound to his/her availability as the number of patients per doctor is increasing day by day. Another existing problem is that a large amount of time and money is spent on making an appointment and visiting a doctor only to get the inference of the medical report. To tackle these problems, many researchers [31] have worked to automate the process of examination of reports by a doctor and work towards making the general public self-capable of analyzing their health reports. The authors of the presented work have tried to achieve the aforementioned targets and even tried to create a platform to increase awareness among the various harmful diseases, their causes, and effects, which must be in the knowledge of the masses. [23] and [11] provide evidence that smartphones and online platforms for healthcare systems have been popular among health professionals as well as patients for a long time now. People are comfortable using these technologies for healthcare analysis.

Early detection of diseases is highly beneficial especially in cases of chronic diseases as it increases the chances of leading a better life with possibilities of cure and longevity of life. According to WHO, by 2020 three quarters of total deaths around the world will be caused by chronic diseases like diabetes, stroke, etc. As it is evident that smartphones and online platforms are becoming a crucial part of people's life, hence, it has the potential to become a powerful tool which will enable a patient or prospective patient to know about the possibility of the existence of a disease in his/her body. With the present boom of the internet worldwide, it is only sensible that we have a medical web application as a solution to the above-stated problems. Our application, AI-DOC, aims at alerting patients about their health conditions and early detection of some chronic and acute diseases alike. This can decrease the mortality rate and save a lot of time and money.

Diseases can be broadly categorized as acute and chronic diseases. Our application includes diseases under both these categories. The broad definition of chronic disease can be stated as the disease which exists for a time of more than 1 year. Such a disease might make a person restricted when it comes to performing usual daily activities and needs continuous medical attention. On the other hand, an acute disease is one that persists for a short period. But this does not denote that the severity of the acute disease will be lesser [20]. AI-DOC predicts diseases like breast cancer, diabetes, and heart diseases that fall under the category of chronic diseases. Liver diseases can be acute as well as chronic. Chronic liver diseases are more common and they occur over a period of time. Malaria, on the other hand, is characterized as an acute disease and with the help of the blood cell picture, AI-DOC can easily classify whether the person is infected by the parasite or not. Pneumonia, an acute respiratory disease, can also be classified by using a lung x-ray.

AI-DOC is a simple solution to detect the possibility of the above-mentioned diseases in a person easily and efficiently. It is an application made using Flask and Gunicorn. It uses the following python libraries: Pandas, NumPy, sci-kit-learn, and Gensim. It also uses the python framework, TensorFlow. It is a user-friendly application and can be used by patients and

doctors alike.

The rest of the paper is organized as follows: Section 2 describes the diseases that have been considered for prediction. The models used are described in detail in section 3. The database used to evaluate the proposed models is presented in section 4. Experimental settings and results for predicting various diseases are explained in section 5. Finally, the conclusion and future scope are mentioned in section 6.

II. Diseases Considered

Diseases like cancer need to be detected at the earliest to increase the chances of survival [3]. The most common cancer is breast cancer among women and according to WHO, 2.1 million women are affected by breast cancer every year. Machine learning and data mining can be used to decrease the quantity of false positive and false negative results while predicting breast cancer [21].

Cardiovascular diseases are common around the world [1] as the lifestyle of people is becoming poorer day by day [27]. Coronary heart diseases come under cardiovascular diseases and it is the blockage of arteries caused by fatty deposits on the internal wall of arteries [2]. This increases the chances of heart attack and hence heart diseases must be detected at the earliest. Various studies have proved machine learning to be quite an efficient tool for the prediction of heart diseases [22], [24].

Diabetes is a metabolic disease that is chronic and is caused due to insufficient production of insulin in the body or when the cells of the body do not acknowledge the insulin or both of the stated reasons [4]. According to WHO, 422 million people worldwide suffer from diabetes. Algorithms like Naïve Bayes, Support Vector Machine (SVM), decision trees, etc. have shown good accuracy in the prediction of diabetes [26], [28], [25].

The liver is a vital organ of the human body and helps in the process of metabolism, the disintegration of red blood cells, and the production of proteins to name a few [18]. Approximately 2 million deaths happen each year worldwide due to liver diseases [10]. Various reasons for liver diseases are smoking, consumption of alcohol, diabetes, and obesity to name a few [32]. Logistic regression, K Nearest Neighbours (KNN), and artificial neural networks can be used to predict liver diseases [15].

Malaria is a disease caused by a parasite that is carried by mosquitoes and has symptoms that include flu-like illness, fever, and chills [33]. According to WHO, in 2018, approximately 228 million cases of malaria occurred worldwide and the number of deaths was 40,5000 [5]. To conquer this problem, researchers have used convolution neural networks to detect parasite-affected cells [12], [14].

Another disease called Pneumonia is caused due to infection of the lungs caused by viruses, bacteria, and fungi [19]. In this, one or both of the lungs swell up and cause discomfort. Around 808,000 children of age below 5 years died due to pneumonia as stated in a report given by WHO in 2017 [8]. Again, convolution neural networks can be used to classify the infected and non-infected lung cells [29]. Even transfer learning can be used to classify the x-ray images of the chest as infected or not [13]. R-CNN has also been used for pneumonia classification [16].

III. Models used

Several machine learning models have been used to study and accurately predict the presence of the aforementioned diseases.

A. Logistic Regression

Logistic Regressions is a supervised classification algorithm, that is used where the outcome variable is categorical. The basic idea is to find a relationship between features and the probability of a particular outcome. To understand logistic regression better, one should first understand linear regression. In linear regression, it is assumed that the data follows a linear function, but in logistic regression models the data uses the sigmoid function, that is,

$$g(z) = 1/(1 + e^{-z}). \tag{1}$$

We know that the linear regression hypothesis is given as

$$h(x) = \theta_0 + \theta_1 x \tag{2}$$

where x is input and θ_0 is bias term and θ_1 is weight of input variable x, and they get updated during training of the model. Hence, when there is a set of independent variables as input, then,

$$h(\theta) = \theta^T X \tag{3}$$

where θ^T is the transpose of the matrix $\begin{bmatrix}\theta_0\\ \theta_1\\ \theta_2\end{bmatrix}$ and X is matrix having all the input features, that is $X = \begin{bmatrix}x_0\\ x_1\\ x_2\end{bmatrix}$. Then in logistic regression, the hypothesis is given by

$$h_\theta(x) = g(\theta^T X) \tag{4}$$

where $g(z)$ is the sigmoid function as discussed above. The loss function, which ascertains how far is the prediction from the original value, is given by

$$J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}[y^i log(h_\theta(x^{(i)})) + (1 - y^{(i)})log(1 - h_\theta(x^{(i)}))] \tag{5}$$

where y is the actual value. Now the aim is to minimize the loss function. This can be done by taking the derivative of the loss function with respect to each weight and changing the weights accordingly until convergence is reached. To minimize loss function we need to run the gradient descent function on each parameter that is,

$$\theta_j = \theta_j - \alpha\frac{\delta J(\theta)}{\delta\theta_j} \tag{6}$$

where α is the learning rate. In this way, the loss function is minimized as the weights are updated. The authors have used logistic regression in the prediction of breast cancer and liver disease.

B. Gradient Boosting

Gradient Boosting is a machine learning algorithm that is used in regression and classification problems. In this, the prediction model is basically in the form of a group of weak prediction models, usually decision trees. Hence, the final prediction is generated by combining the predictions of the multiple decision trees. We can understand gradient boosting easily if we know how the AdaBoost algorithm works. In AdaBoost, a decision tree is trained by first assigning equal weight to each observation. When the first tree has been evaluated, we then increase or decrease the weights of observations according to the level of difficulty in classifying them. Hence, the second tree is essentially made using these weights. The aim is to improve the predictions of the first tree. Now, the new model will be the sum of first two trees. After computing the classification error of the two tree ensemble model, the third tree is made and, in this way, the process is repeated for a specific number of iterations. The final prediction of the ensemble model is the weighted sum of the predictions made by the fixed number of previous decision tree models. In gradient boosting, one can say that the models are trained in an additive, gradual and sequential manner. In the proposed work, gradient boosting is used in the prediction of diabetes.

C. K – Nearest Neighbors

This machine-learning algorithm is a supervised machine learning algorithm that is quite simple and is used in both classification and regression problems. In this, a data point is given a value based on how close it is to the points in the training set. First, the value of K is defined, which is an integer. For a given data point, its distance from the K nearest points in the training set is then calculated. The distance is taken as Euclidean's distance given by,

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{7}$$

The Manhattan distance can also be used instead of the Euclidean's Distance. The class of the new data point is assigned according to the most common neighbor from among the nearest K points. The authors have used K- nearest neighbor models in predicting heart disease.

D. Convolution Neural Network

In the prediction of malaria and pneumonia, the authors have used the dataset in the form of images of blood cells and x-ray images of the chest. The best method for image classification is a specialized neural network called the convolution neural network. If there is a need to process data having the input shape of a 2D matrix, for example, an image, then a convolution neural network is the best and most useful tool. Convolution is a mathematical operator and here we perform convolution on the image x which is a 2D image and a filter w which is useful in identifying various features of the image like edges, vertical and horizontal lines, etc. The output we get after performing convolution is called a feature map. The feature map $s(i,j)$ is given as,

$$\begin{aligned} s(i,j) &= x(i,j) * w(i,j) \\ &= \sum_m \sum_n x(m,n) \cdot w(i-m, j-n) \end{aligned} \tag{8}$$

Different filters are convolved with the 2D array and thus, different features maps are obtained which are then passed through the ReLu function, given by,

$$R(z) = max(0, z) \tag{9}$$

to remove all the negative values. A sigmoid function can also be used but the ReLu function is preferred over sigmoid. By using multiple convolution layers, that is, by performing multiple convolutions on the feature maps which are extracted, a deep network can be made which will help in extracting high-level features from the input images. In the end, a fully connected layer is present where classification takes place.

IV. Dataset Description

1) UCI machine learning repository [9]: To acquire the medical datasets for the prediction of breast cancer, heart, diabetes, and liver diseases, the authors have used the UCI ML repository which is open source, containing a total of 559 datasets. This repository was initiated by David Aha and his fellow graduate students and since then it has been extensively used by students, researchers, and educators across the world as a major source of datasets. It is a collection of various datasets, data generators, and domain theories that is helping the machine learning community to grow through deeper analysis of machine learning algorithms.
2) National Library of Medicine [7]: For malaria predictions, the authors have used the malaria datasets from the national library of medicine. The repository contains images of segmented cells on a slide. The slide was smeared with a thin layer of blood. The dataset is a part of the Malaria Screener research activity. To help researchers in resource-constrained areas, the images of blood smeared slides have been manually annotated by slide readers who are experts in their field of work. A level set–based algorithm was then applied to detect the red blood cells and segment them.
3) Mendeley Data [6]: Chest X-ray images are required to predict pneumonia. Thus the authors have used the Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for classification dataset [17]. The dataset contains validated images of OCT and x-rays of the chest.

V. Results and Discussion

A. Breat Cancer

To predict the existence of breast cancer tumors, the authors have used a Logistic Regression model. The various correlations are depicted in figure 6. The performance matrix for breast cancer prediction is given in table I.

TABLE I
PERFORMANCE MATRIX FOR PREDICTING BREAST CANCER.

Accuracy	Precision	Recall	F1 score
0.986	1.0	0.957	0.978

B. Cardiovascular Diseases

Several machine learning models were used to predict cardiovascular diseases which are Random Forest, Logistic Regression, K Nearest Neighbours, Naive Bayes, and Decision Tree. KNN proved to be the most appropriate model as it predicted the results with the highest accuracy as seen in figure 1. Even though KNN and Random Forest have nearly equal accuracy, KNN is preferred because it is a lightweight model and is easy to deploy. The ROC Characteristics are shown in figure 2.

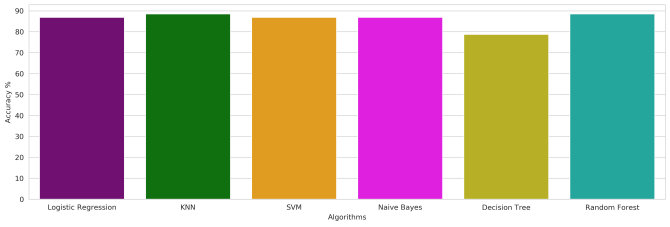

Fig. 1. Comparison of Various Machine Learning Models for predicting Cardiovascular Diseases

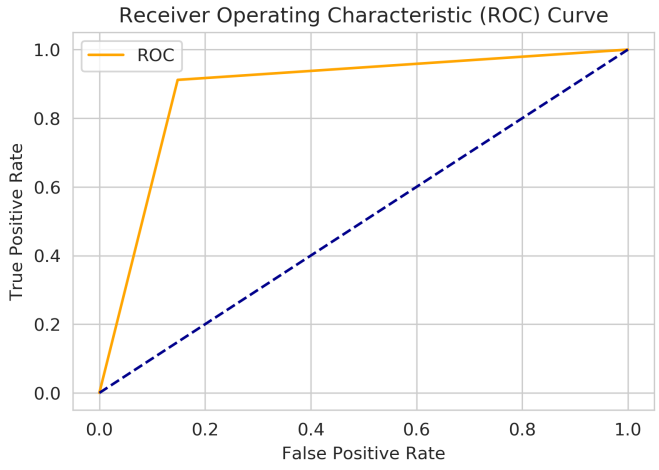

Fig. 2. ROC Characteristics for KNN Model for Cardiovascular Diseases.

C. Diabetes

For predicting various types of Diabetes, the authors have chosen LightGBM and KNN models. The performance report is shown in figure 3.

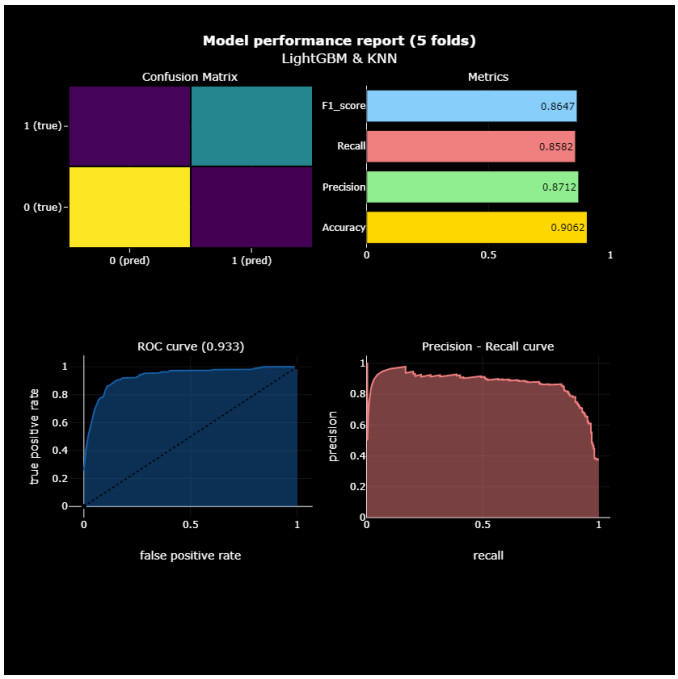

Fig. 3. Performance Report for LightGBM and KNN for Diabetes Prediction.

D. Liver Disease

A Logistic Regression model is used for the prediction of Liver Disease. The performance matrix is shown in table II

TABLE II
PERFORMANCE MATRIX FOR PREDICTING LIVER DISEASE.

Accuracy	Precision	Recall	F1 score
0.886	0.782	0.843	0.811

E. Malaria

The authors have used image dataset to predict Malaria using Convolutional Neaural Networks. The accuracy and loss characteristics are shown in figure 4.

F. Pneumonia

Convolutional Neural Networks are used for predicting Pneumonia using images. The accuracy and loss characteristics are shown in figure 5.

VI. Conclusion

In the proposed work the authors were able to use different machine learning algorithms to predict different diseases with satisfying accuracy. Humongous amounts of data have been analyzed closely to select the appropriate model. In this way, the authors were able to include all the diseases under one cap in the form of a web application. The user interface of

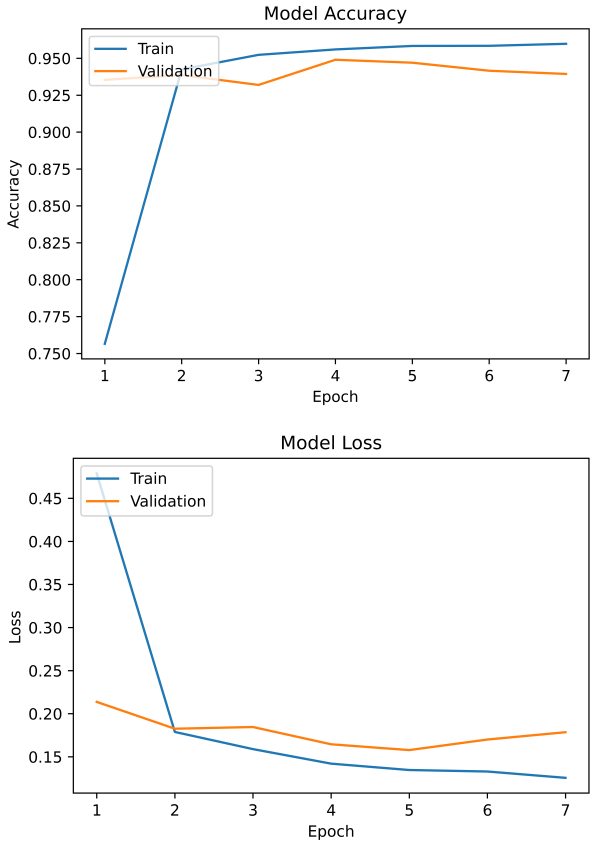

Fig. 4. Accuracy and Loss Characteristics for Malaria Prediction

the application is simple. The patients can enter the medical parameters, predict the possibility of a disease and then make an appointment with the doctor according to the need. Hence, we were able to create a platform that can help the patients initially understand their medical reports without the assistance of a doctor. The presented work by no means suggests the patients rely on machine-generated prediction only. The motive of AI-DOC is to provide early assistance to the patients in understanding their medical health reports and reduce the time and money spent in visiting a doctor for initial checkups. This in turn helps in reducing the doctors' stress to tackle a large number of patients on a daily basis, thus increasing their efficiency as a doctor. Since the application involves the use of a wide set of personal medical information, the authors aim to include a login facility to maintain privacy.

References

[1] Cardiovascular disease - world health organisation. https://www.who.int/cardiovascular_diseases/about_cvd/en/, .

[2] Coronary heart disease. https://www.nhp.gov.in/disease/cardio-vascular/heart/coronary-heart-disease, .

[3] Promoting cancer early diagnosis - world health organisation. https://www.who.int/cancer/prevention/diagnosis-screening/en/.

[4] Diabetes mellitus. https://www.nhp.gov.in/disease/digestive/pancreas/diabetes-mellitus.

[5] Malaria - world heath organisation. https://www.who.int/news-room/fact-sheets/detail/malaria.

[6] Labeled optical coherence tomography (oct) and chest x-ray images for classification - mendeley data. https://data.mendeley.com/datasets/rscbjbr9sj/2.

[7] Malaria dataset, national library of medicine. https://lhncbc.nlm.nih.gov/LHC-publications/pubs/MalariaDatasets.html.

[8] Pneumonia - world heath organisation. https://www.who.int/health-topics/pneumonia/#tab=tab_1.

[9] Uc irvine machine learning repository. https://archive.ics.uci.edu/ml/index.php.

[10] S. K. Asrani, H. Devarbhavi, J. Eaton, and P. S. Kamath. Burden of liver diseases in the world. *Journal of hepatology*, 70(1):151–171, 2019.

[11] M. M. Baig, H. GholamHosseini, and M. J. Connolly. Mobile healthcare applications: system design review, critical issues and challenges. *Australasian physical & engineering sciences in medicine*, 38(1):23–38, 2015.

[12] K. Fuhad, J. F. Tuba, M. Sarker, R. Ali, S. Momen, N. Mohammed, and T. Rahman. Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application. *Diagnostics*, 10 (5):329, 2020.

[13] M. F. Hashmi, S. Katiyar, A. G. Keskar, N. D. Bokde, and Z. W. Geem. Efficient pneumonia detection in chest xray images using deep transfer learning. *Diagnostics*, 10(6):417, 2020.

[14] J. Hung and A. Carpenter. Applying faster r-cnn for object detection on malaria images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 56–61, 2017.

[15] J. Jacob, J. C. Mathew, J. Mathew, and E. Issac. Diagnosis of liver disease using machine learning techniques. *Int. Res. J. Eng. Technol.(IRJET)*, 5(4):4011–4014, 2018.

[16] A. K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, and J. J. Rodrigues. Identifying pneumonia in chest x-rays: A deep learning approach. *Measurement*, 145:511–518, 2019.

[17] D. Kermany, K. Zhang, and M. Goldbaum. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2, 2018.

[18] Y. Kumar and G. Sahoo. Prediction of different types of liver diseases using rule based classification model. *Technology and Health Care*, 21(5):417–432, 2013.

[19] L. A. Mandell and M. S. Niederman. Aspiration pneumonia. *New England Journal of Medicine*, 380(7):651–663, 2019.

[20] A. J. Marcano-Reik. *Acute Disease*. Springer New York, New York, NY, 2013. ISBN 978-1-4419-1005-9. doi: 10.1007/978-1-4419-1005-9_1202. URL https://doi.org/10.1007/978-1-4419-1005-9_1202.

[21] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake. Analysis of breast cancer detection using different machine learning techniques. In *International Conference on Data Mining and Big Data*, pages 108–

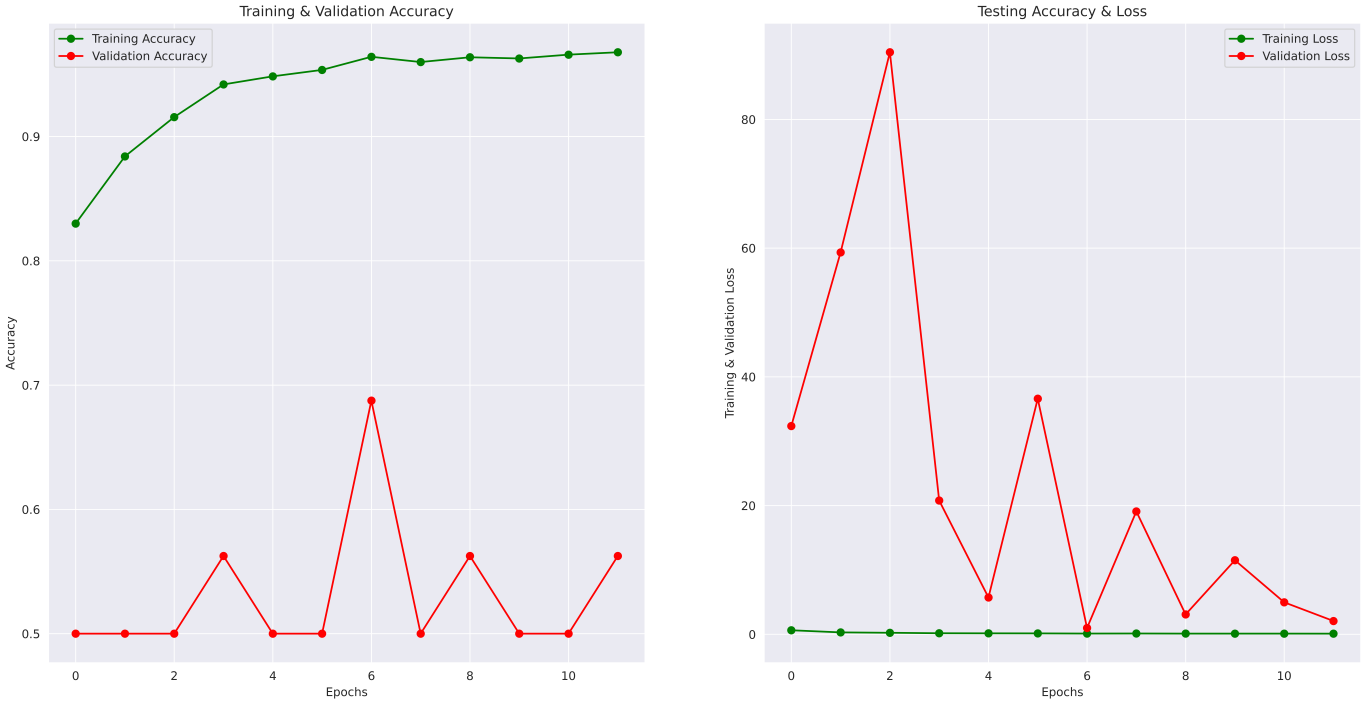

Fig. 5. Accuracy and Loss Characteristics for Pneumonia.

117. Springer, 2020.

[22] S. Mohan, C. Thirumalai, and G. Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7:81542–81554, 2019.

[23] A. S. M. Mosa, I. Yoo, and L. Sheets. A systematic review of healthcare applications for smartphones. *BMC medical informatics and decision making*, 12(1): 67, 2012.

[24] V. Ramalingam, A. Dandapath, and M. K. Raja. Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, 7(2.8):684–687, 2018.

[25] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah. Prediction of diabetes using machine learning algorithms in healthcare. In *2018 24th International Conference on Automation and Computing (ICAC)*, pages 1–6. IEEE, 2018.

[26] D. Sisodia and D. S. Sisodia. Prediction of diabetes using classification algorithms. *Procedia computer science*, 132:1578–1585, 2018.

[27] J. Slavíček, O. Kittnar, G. E. Fraser, E. Medová, J. Konečná, R. Žižka, A. Dohnalová, and V. Novák. Lifestyle decreases risk factors for cardiovascular diseases. *Central European journal of public health*, 16(4): 161, 2008.

[28] N. Sneha and T. Gangil. Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big data*, 6(1):13, 2019.

[29] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong. An efficient deep learning approach to pneumonia classification in healthcare. *Journal of healthcare engineering*, 2019, 2019.

[30] K. M. Sutcliffe, E. Lewton, and M. M. Rosenthal. Communication failures: an insidious contributor to medical mishaps. *Academic medicine*, 79(2):186–194, 2004.

[31] U. Varshney. Pervasive healthcare: applications, challenges and wireless solutions. *Communications of the Association for Information Systems*, 16(1):3, 2005.

[32] S. Vijayarani and S. Dhayanand. Liver disease prediction using svm and naïve bayes algorithms. *International Journal of Science, Engineering and Technology Research (IJSETR)*, 4(4):816–820, 2015.

[33] N. J. White. Anaemia and malaria. *Malaria Journal*, 17 (1):1–17, 2018.

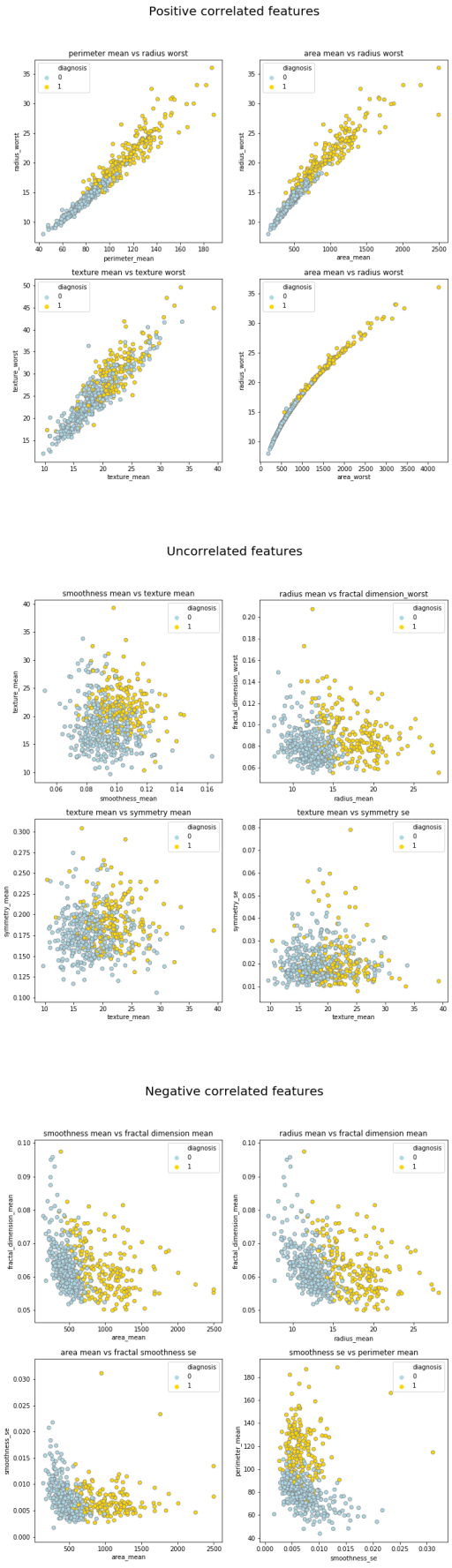

Fig. 6. Correlation of Various Features for Breast Cancer.