21BIO211 INTELLIGENCE OF BIOLOGICAL SYSTEMS

Web Interface For Predicting Multiple Diseases

PREPARED BY

CH.EN.U4AIE21107	**APARNA S**
CH.EN.U4AIE21113	**ROHAN G V**
CH.EN.U4AIE21115	**HARSHITHA M**
CH.EN.U4AIE21122	**KORRAYI SAITEJA**
CH.EN.U4AIE21125	**M SRAVANI**

ACKNOWLEDGEMENT

We offer our sincere pranam at the lotus feet of Universal guru, **MATA AMRITANANDAMAYI** who blessed us with her grace to make this successful major project.

We express our genuine sense of greatefulness to **Dr. V Jayakumar** Principal of Amrita School of Engineering, Chennai and **Dr. Manikandan I B** Director, Amrita School of Engineering for their kind support. We are grateful to our guide Dr. I R Oviya-Sciences-ASE-Chennai Campus for her acute interest, advice, constant encouragement during every stage of our work to bring this work into success.

We are also appreciative to our guide Dr. Prasanna Kumar R, Chairperson of Computer Science (Artificial Intelligence), ASE Chennai for his guidance. We are also thankful to Dr. Sasikala, Class Advisor of AIE B, who helped us with this project. We are also thankful to other staff members of Department of sciences for their valuable help, all our classmates who have always been a source of strength for always being there and extending their valuable helps to the successful completion of this work.

Abstract

Prediction of diseases plays a crucial role in clinical diagnosis of the patient's treatment based on the symptoms. Most of the computational tools are designed for mobile applications and web applications in the healthcare field for detection or prediction purposes. These predictions can limit the cost that we spend on the test, scanning, etc. Moreover, this web interface is an alternate solution that can be used in emergency situations which is readily available on the web page for immediate prediction of disease with maximum accuracy. This web interface reduces the level of risk by predicting diseases with immediate results to speed up the recovery process.

In this paper, we present an ML-based user-friendly web interface "Swasthgarud" which is designed based on SVM (Support vector Machine), XG Boost, Random Forest, CNN algorithms, and Gradient for the prediction of diseases like Heart disease, Pneumonia, Malaria, and diabetes in women by taking input image datasets of X-rays, blood samples.

***Keywords*:** Prediction, Stream lit, web interface, Accuracy, Machine learning (ML), Convolutional Neural Networks (CNN), Random Forest, Support Vector Machine (SVM) Heart disease, X-ray, Pneumonia, Malaria, diabetes, Blood samples

1. Introduction

More than preventing and curing, predicting a disease accurately can drastically improve healthcare outcomes. And this can be achieved with Machine Learning. It has the potential for a wide range of applications in healthcare, including predictive analysis, drug discovery, improving the quality of patient care, increasing healthcare efficiency, accurate diagnosis, and so on. Diseases such as pneumonia, diabetes, heart disease, and malaria can be a serious threat if they are not predicted accurately. Therefore, we created a web interface that uses machine learning algorithms to make predictions on various diseases like Pneumonia, Malaria, Diabetes, and heart disease. These machine learning models are deployed with Stream lit – an open-source Python framework to build web apps.

2. Problem statement

This problem includes designing a user-friendly web interface to predict the disease with a maximum accuracy rate by using machine learning algorithms. Isn't it very difficult for doctors to analyze X-ray images and MRI images of patients every time? Doctors will face many challenges while analyzing the images of X-ray and MRI images. Some of the challenges are

Quality of Image: Image quality is the most crucial task for predicting accurate results. It decides the predictions based on the image quality.

Limitation of time: Doctors often will have limited time to analyze the X-ray and MRI images which leads to errors.

3. Proposed Solution

We are developing a website for predicting multiple diseases. In this website it is possible to predict more than one disease at a time. So, the user doesn't need to change the websites to predict the diseases. We are taking four diseases that are Pneumonia, Heart diseases, Diabetes, Malaria. To implement multiple disease predictors, we are going to use machine learning algorithms like random Forest, Convolutional Neural Network, Support Vector Machine and a python open-source framework- Stream lit. When the user is accessing the website, the user has to send the parameters of the disease along with the disease name. Streamlit enables users to develop an attractive user web interface. It helps in building machine learning and data science applications.

4. Literature Survey

- **Web-Application development tasks of prediction in medical domain [1]:** This paper was published in the year 2018. The paper deals with clinical disease development analysis and hence creating a web application with RStudio- shiny library.
- **Web application based disease predictor and recommender system [2]:** This paper was published in the year 2021. This paper deals with ML based prediction on the datasets retrieved from the National Center of Disease Control.

  Ml algorithms: KNN, Random Forest, Ensemble voting algorithm, Naïve Bayes Classifier.

  Dataset: Dataset acquired from the NCDC.
- **A real-time patient monitoring system for heart disease prediction using a random forest algorithm [3]:** This paper was published in the year 2015. deals with patient monitoring systems that are meshed with the patient monitoring system. Acts as a decision making that provides the early detection latter reduced the time before the treatment of heart disease. With the means of different wireless technologies, the forward alarms are generated to the relevant curator.

  Algorithm: Random Forest
- **Development of a Python classification web interface for independent dataset [4]:** This paper was published in the year 2022. This paper mainly deals with developing a web interface for classification analysis. Sklearn and dash are used in building this web interface. Algorithms: Random Forest, naïve Bayes algorithm, LightGBM, Support

Vector Machine, XGBoost and AdaBoost.
Dataset: Cancer dataset.

- **A Web based chest X-Ray Disease Predictor [5]**: This paper was published in the year 2020. For diagnosis, the image data is provided and gets classified. The web interface runs locally without sending data to the device. Scalabitlity is extended up to million users. **Dataset:** 7 largest radiological public datasets.
- **F-SNP computationally predicted SNPs for disease [6]:** This paper was published in the year 2007. This system combines databases and bioinformatics tools. Further finds the impact of Single Nucleotide Polymorphism on humans.
- **Dataset:** F-SNP dataset
- **Web-application for online pneumonia risk prediction system [7]:** This paper was published in the year 2018. The existing models deal with predicting the stroke risk factor in the patient and reduce the chances of ICU admission of patients.
- **Datasets:** samples from 20 pneumonia patients and 20 healthy subjects between 20 to 30 years old ages.
- **Classification of diabetes with Random Forest- feature selection algorithm [8]:** This paper was published in November 2019. This paper uses algorithms like Support vector machine, Naïve Bayes algorithm, Random Forest, Linear and Logistic Regression for diabetes prediction. The accuracy of the Random Forest to predict diabetes is 85.6 and that of with feature selection, it is obtained over 92 percent.
Datasets: The pima Indian diabetes dataset. This contains 768 instances and 8 features including age, glucose level, blood pressure, BMI.

5. System Analysis

5.1 Functional Requirement

The system allows the user to upload the necessary image as the input according to the machine learning model and the predicted output should be displayed.

5.2 Non- Functional Requirement

The website should be:

- User-friendly
- Attested
- Scalable
- Compatible

6. Web Interface

Figure 6.1: Landing Page of the web interface

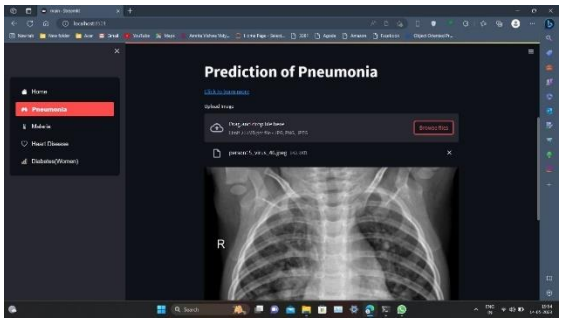

Figure 6.2: Prediction of Pneumonia

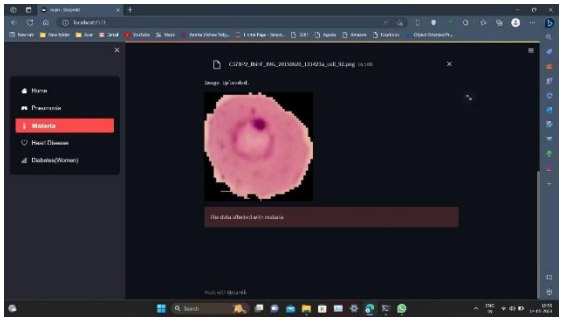

Figure 6.3: Prediction of Malaria

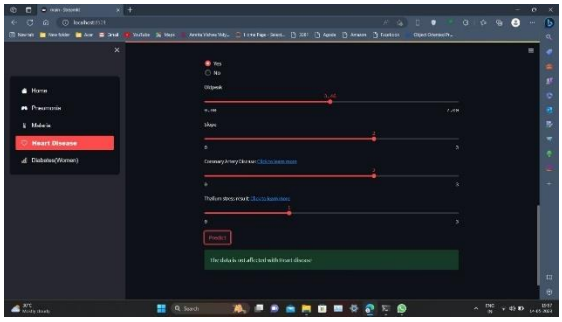

Figure 6.4: Prediction of Heart Disease

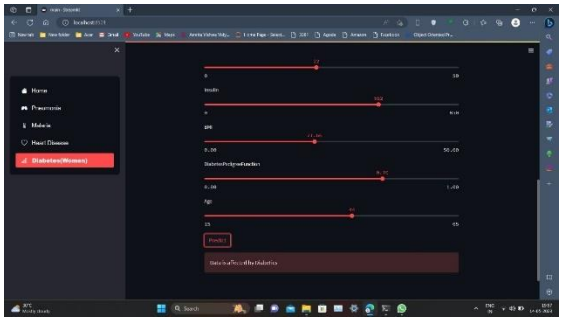

Figure 6.5: Prediction of Diabetes

7. Methodology

The proposed system "Swasthgarud" uses machine learning algorithms like CNN and Random Forest, etc to predict the accurate results. Our work includes diseases like Pneumonia, Diabetes, Heart disease, and Malaria. Image datasets of red blood cells are considered for Malaria, X-ray image datasets are fed as input for pneumonia, and image datasets of ECG, and BP are fed as input for heart diseases. Datasets are retrieved from Kaggle. Deployment of the model into the web interface is deployed using "Stream lit". "Adobe XD" is used for prototyping our model. Below is the table for the datasets of the respective disease table 1.1.

The design and implementation of the model include data collection, data preprocessing, data labeling, splitting data, selecting the model architecture, training the model, evaluating the model, recall, testing the model, deploying the model into the web interface, and the last step is converting the model into the web interface. Figure 1.2 shows the workflow of the system.

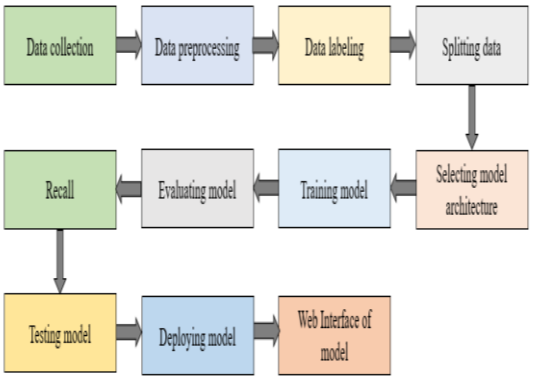

Figure 7.1: Workflow of the model

7.1 Prediction of Pneumonia

About 7% of people are getting infected by Pneumonia. It is caused by bacteria, fungi, or viruses in the lungs. The formation of mucus (sticky layer) in the lungs causes symptoms like cough, fever, breathing problems, and chest pain. There are many other computational tools developed for the prediction of pneumonia. This disease can be predicted through Chest X-ray, Blood tests, CT scan, Sputum culture, and physical examination. In our model, we train the datasets of X-ray images of pneumonia using CNN (convolutional neural network). This model helps to predict the risk factor in patients.

7.2 Prediction of Heart Disease

In this modern era, there is a rapid increase in heart disease every year without age restriction. Therefore, it has become one of the major problems for prediction and finding the best treatment for the disease to reduce the chance of risk in the patient. The major factors that develop the maximum chance of heart disease are Blood pressure, hereditary, maximum percentage of cholesterol, and obesity. Based on the prediction results further process the risk factor will be estimated and helps to suggest the best-personalized plan for the patient. This system limits the cost which will be spent on tests, scanning, and other diagnostic tests. Heart disease can be predicted through various diagnostic tests such as Blood tests, ECG, cardiac catheterization, and stress tests. Detection and prediction of heart disease is implemented on Random Forest which gives the best accurate results out of all algorithms for image datasets of ECG and cholesterol.

7.3 Prediction of Malaria

Based on the previous survey, WHO (world health organization) estimates around 229 million patients are affected by malaria. This is considered as one of the most harmful diseases due to inadequate treatment and medical care support for the patients. Through image processing blood smear samples are analysed for the prediction of the disease. The detection of this disease is not only helpful for the user but also for the lab technicians to analyse the immediate results as the samples of the blood smear images of the patient will not be readily available for the users. The major factors that affect malaria are environmental factors which cause the rapid growth of mosquitoes which causes Malaria. Convolutional Neural Networks (CNN) predicts the results most accurately in analyzing the datasets.

7.4 Prediction of Diabetes

Diabetes is considered as one of the most dangerous diseases because it increases the sugar level in the blood, which develops a high chance of risk factors. It is basically of two types type 1 and type 2. Irrespective of age and gender six in ten get infected with diabetes. In this paper, we mainly focus on predicting diabetes in women. Women easily get infected with diseases in comparison with men because of various factors some of them are family history of diabetes, pregnancy, obesity,

lack of physical activity in some women, and high blood pressure. In this model diabetes in women will be predicted based on the answers provided in the queries by the user in the web interface. This reduces the risk of developing diabetes by knowing the symptoms well in advance. Women easily get infected with diseases in comparison with men because of various factors some of them are family history of diabetes, pregnancy, obesity, lack of physical activity in some women, and high blood pressure.
This model is developed based on the machine learning algorithms like Random Forest, SVM, Decision Tree, Gradient Boosting, XGBoost, Logistic Regression by retrieving BMI (Body Mass Index), Insulin, Glucose level, BP (Blood Pressure) datasets from UCI to get the best results with maximum accuracy.

8. Algorithms

8.1 Logistic Regression

Step_1: Import the dataset and segregate the independent and dependent variables

Step _2: Then apply logistic regression to the dataset.

Formula: F(x) = $1/1 + e^{-(\beta_0+\beta_1 x)}$

β_0 is the slope

β_1 is the y-intercept

X is independent variable.

Step 3: Predicting the values.

8.2 KNN algorithm

Step_1: First fix value of K which means number of neighbors.

Step_2: Then calculate the distance between the points by Euclidean distance. Let's understand with an example.

Example: A(1,2) and B(3,4)

Distance =$\sqrt{(3-1)^2+(4-2)^2}$

Step_3: Then calculate the Euclidean distance and group them to nearest neighbors.

Step_4: Calculate the number of k-neighbours in each Category.

Step_5: Assign the new data points to category of k neighbours which has maximum numbers of points with minimum distance.

8.3 Decision Tree

Step_1: Initialize the dataset and define the attributes

Step_2: Calculate entropy and the information gain of every attribute that present in the dataset

Formula of Entropy: E(T, X) = $\sum_{\mathbb{C}\in x} p(c)E(c)$

$$E(C) = \sum_{i=1}^{C} -\mathbb{P}_i \log \mathbb{P}_i$$

P_i = probability of an event of that attribute

Information Gain = Entropy(T) – Entropy(T,X)

Step_3: Select an attribute from the dataset which has high information gain or lower entropy as the root node of tree.

Step_4: Split the Dataset as selected attribute to obtain the subset of the data.

Step_5: Repeat the above steps until every attribute is selected

8.4 Random Forest:

Step_1: First select the k random attributes from the dataset.

Step_2: Construct a decision tree for each selected attribute from dataset.

Step_3: Based on result produced by each decision tree voting will takes places.

Step_4: Considering the maximum voting will be selected for the predicting the data.

8.5 Gradient Boosting

Step_1: Initialize the dataset and model with the constant value and apply the loss function.

Step_2: Calculate the Gradient and build a new decision tree based on the gradient

Step_3: Then the model is ready to predict the values.

8.6 XGBoost

Step_1: First initialize the dataset and a single leaf tree.

Step_2: Calculate the average target attributes to predict and residuals make use of the loss function.

Step_3: Calculating the similarity score

Similarity score = Gradient *(Gradient2/Hessian +λ)

Gradient2 is squared sum of residuals

λ is the regularization hyperparameter

Step_4: Calculate the information gain.

Information Gain = left Similarity + Right Similarity – Similarity of roots

Step_5: Create a tree of required length by selectively removing unwanted branches and regulating with regulation hyperparameter.

Step_6: Then predict the residuals with help of decision tree and calculate the new set of residuals

New Residuals = old + $\rho\Sigma$Predicted

Step_7: Repeat the process for all tress to achieve the final output.

8.7 ConvNet

Step_1: Muster the suitable image for predicting the disease.

Step_2: Pre-process the image dataset.

Step_3: Break down the data into two sets: training data and the Testing data.

Step_4: After defining the architecture of the CNN, train the model using the optimizer and loss function.

Step_5: The model should be evaluated with an F1 score.

Step_6: To predict, the new image should pass onto CNN which in turn returns the output based on the interpretation of the probability score.

8.8 SVM image classifier

Step_1: Muster the images for testing and training the SVM classifier.

Step_2: Pre-process the image dataset.

Step_3: The model should be trained with the training set and should be evaluated with an F1 score.

Step_4: To predict, the features of the new image is pulled out and passed through SVM model.

9. Retrieved Datasets

9.1 Pneumonia Dataset

To predict pneumonia, we used “Chest X-ray images (Pneumonia)” from Kaggle. This dataset consists of 5863 X-ray images (JPEG) and has two categories: Pneumonia and Normal. These images were collected from paediatric patients of one to five years old.

9.2 Malaria Dataset

To predict malaria, we used "Malaria Screener datasheet” from National Library of Medicine. This dataset consists of 27,558 images (JPEG) and has two categories: Infected and Uninfected.

9.3 Diabetes Dataset

To predict diabetes, we used “Diabetes dataset” from University of California, Irvine Machine Learning. It has 8 predictor features with one target feature. Pregnancies, Skin thickness, BMI, Age, Glucose level, Pedigree are some of the attributes used.

9.4 Heart Disease Dataset

To predict heart disease, we used “heart disease dataset” from University of California, Irvine Machine Learning. It has 76 attributes including predicted attribute.

10. Result

Accuracy for each predicated disease is listed below:

10.1 Heart Disease

S.No	Algorithm	Testing accuracy (%)	Training accuracy (%)
1	Logistic Regression	85.49	88.31
2	KNN	87.86	86.03
3	SVM	71.04	71.45
4	Decision tree	100	96
5	Random forest	98.88	98.05
6	Gradient Boosting	98.60	97.72
7	XGBoost	97.907	97.07

Table 10.1: Obtained accuracy for heart disease.

10.2 Diabetes

S.No	Algorithm Name	Testing accuracy(%)	Training accuracy(%)
1	Logistic Regression	84.02	88.15
2	KNearestNeighbors	86.6	83.33
3	SVM	89.47	84.42
4	Decision Tree	86.54	89.47
5	Random Forest	98.30	92.25
6	Gradient Boosting	100	89.03
7	XGBoost	98.49	97.5

Table 10.2: Obtained accuracy for diabetes.

10.3 Malaria

S. No	Algorithm	Accuracy
1.	CNN	90.05
2.	SVM	50

Table 10.3: Obtained accuracy for predicting Malaria.

10.4 Pneumonia

S. No	Algorithm	Accuracy
1	CNN	95.2
2	SVM	50

Table 10.4: Obtained accuracy for predicting Pneumonia.

11. Conclusion

The main intention of this project is to predict multiple diseases such as Pneumonia, Malaria, heart disease and diabetes. In addition to that, we aim to achieve maximum accuracy in predicting these diseases. The trained machine learning models are deployed in Stream lit – a python framework to make it as a user-friendly web interface. The machine learning algorithms used are Convnets, Decision Tree Logistic Regression, Support Vector Machine, K- Nearest Neighbour, and Random Forest.

12. Future Scope

Subsequently, we can extend this project by appending more diseases that can be predicted by ML and Deep Learning algorithms. Also, we can improve scalability of the web interface.

13. References

[1] M. Mazorchuck, V. Dobriak and D. Chumachenko, "Web-Application Development for Tasks of Prediction in Medical Domain," *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, Ukraine, 2018, pp. 5-8, doi: 10.1109/STC-CSIT.2018.8526684.

[2] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Application of dimensionality reduction in recommender system-a case study. Minnesota Univ Minneapolis Dept of Computer Science.

[3] Sreejith, S., Rahul, S., & Jisha, R. C. (2016). A real time patient monitoring system for heart disease prediction using random forest algorithm. In Advances in Signal Processing and Intelligent Recognition Systems: Proceedings of Second International Symposium on Signal Processing and Intelligent Recognition Systems (SIRS-2015) December 16-19, 2015, Trivandrum, India (pp. 485-500). Springer International Publishing.

[4] Cicek, İ. B., İlhami, S. E. L., YAĞIN, F. H., & Colak, C. (2022). Development of a Python-Based Classification Web Interface for Independent Datasets. Balkan Journal of Electrical and Computer Engineering, 10(1), 91-96.

[5] Cohen, J. P., Bertin, P., & Frappier, V. (2019). Chester: A web delivered locally computed chest x-ray disease prediction system. arXiv preprint arXiv:1901.11210.

[6] Phil Hyoun Lee, Hagit Shatkay, F-SNP: computationally predicted functional SNPs for disease association studies, *Nucleic Acids Research*,

Volume 36, Issue suppl_1, 1 January 2008, Pages D820–D824

[7] W. N. W. M. Afifi, I. F. Warsito, M. Sayahkarajy and E. Supriyanto, "The development of an online pneumonia risk prediction system," *2017 International Conference on Robotics, Automation and Sciences (ICORAS)*, Melaka, Malaysia, 2017, pp. 1-5, doi: 10.1109/ICORAS.2017.8308063.

[8] Chari, K. K., M. Chinna Babu, and S. Kodati. "Classification of diabetes using random forest with feature selection algorithm." Int. J. Innov. Technol. Explor. Eng 9.1 (2019): 1295-1300.