*Human Multiple Disease Prediction*

Rangampet Vikas   
Computer Science and Engineering*.*   
Hyderabad Institute of Technology and ManagementHyderabad, India

Peddi Akash  
Computer Science and Engineering  
Hyderabad Institute of Technology and Management Hyderabad, India  
[Padala](mailto:vishaljavvaji650@gmail.comPadala) Raviteja  
Computer Science and Engineering  
Hyderabad Institute of Technology and ManagementHyderabad, India  
ravitejapadala9@gmail.com

Veldhandi Sai KrishnaComputer Science and Engineering  
Hyderabad Institute of Technology and ManagementHyderabad, India

Dr. Padmaja Pulicherla   
Computer Science and Engineering  
Hyderabad Institute of Technology and ManagementHyderabad, India  
[padmaja.j2ee@gmail.com](mailto:padmaja.j2ee@gmail.com)

padmajap.cse@hitam.org

***Abstract: Many of the existing machine learning models for health care analysis are concentrating on one disease per analysis. Like one analysis for diabetes analysis, one for Heart analysis, one for lung diseases, one for kidney analysis like that. There is no common system where one analysis can perform more than one disease prediction. In this proposing system which is used to predict multiple diseases by using Flask. In this system used to analyze Diabetes, Cancer, Heart, Kidney, Liver, Malaria, Stroke, Pneumonia, the accuracy of medical disease prediction has been continuously improved, and the performance in all aspects has also been significantly improved. It aims to clarify the effectiveness of machine learning in disease prediction and demonstrates the high correlation between machine learning and the medical field in future development. The performance of the proposed system is evaluated on the scales of accuracy. The results reveal the effectiveness of our proposed methodology in predicting multiple diseases in comparison to other benchmark methods. The disease prediction is accomplished based on the features extracted from the raw dataset. To implement multiple disease analysis used machine learning algorithms and Flask Python pickling is used to save the model behavior. The importance of this analysis is to analyze the maximum diseases, so that to monitor the patient's condition and warns the patients in advance to decrease mortality ratio.***

***Keywords: Machine learning, Flask, Decision Tree, Support Vector Machine (SVM), Random Forest, LR, KNN.***

# INTRODUCTION

In this digital world, data is an asset, and enormous data was generated in all the fields. Data in the healthcare industry consists of all the information related to patients. Here a general architecture has been proposed for predicting the disease in the healthcare industry. Many of the existing models are concentrating on one disease per analysis. Like one analysis for diabetes analysis, one for cancer analysis, one for skin diseases like that. There is no common system present that can analyze more than one disease at a time. Thus, we are concentrating on providing immediate and accurate disease predictions to the users about the symptoms they enter along with the disease predicted. So, we are proposing a system which used to predict multiple diseases. In this system, we are going to analyze Diabetes, Cancer, Heart, Kidney, Liver, Malaria, Stroke and Pneumonia disease analysis. Later many more diseases can be included In multiple disease prediction, it is possible to predict more than one disease. The distinctive feature extraction ways of deep learning ways will still play a crucial role in future medical analysis. Advancement in analytical models, handiness of GPU hardware, and cloud infrastructure began to play a important role in attention practices and analysis. it's various tools and techniques to archive, manage, analyse, and predict giant volumes of structures, unstructured and semi structured knowledge. knowledge Science plays an important role in medical fields with higher support for identification and cure for the diseases. during this project, we have a tendency to aiming to predict multiple diseases like cancer, Diabetes, heart, liver, kidney, malaria, respiratory illness by victimization totally different In this digital world, data is an asset, and enormous data was generated in all the fields. Many of the existing models are concentrating on one disease per analysis. Our Model will invoke the corresponding model and return the status of the patient. Our basic idea is to develop a system which will predict and give the details of the disease predicted along with its severity which as symptoms are given as input by the user. The system will compare the symptoms with the datasets provided in the database. If the symptom matches the datasets, then it should ask other relevant symptoms specifying the name of the symptom. If not, the symptom entered should be notified as the wrong symptom. After this a prompt will come up asking whether you want to still save the symptom in the database. If you click on yes, it will be saved in the database, if not it will go to the recycle bin. The main feature will be the machine learning, in which we will be using algorithms such as Naive Bayes Algorithm, K- Nearest Algorithm, Decision Tree Algorithm, Random Forest Algorithm and Support Vector Machine, which will predict accurate disease and also, will find which algorithm gives a faster and efficient result by comparatively- comparing The importance of this system analysis is that while analyzing the diseases all the parameters which cause the disease are included so it is possible to detect the disease efficiently and more accurately. The final model's behavior will be saved as a python pickle file.

# LITERATURE SURVEY

According to the paper, 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 forecast whether 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. Here they used mainly 4 main algorithms Decision Tree, Naïve Bayes, and SVM algorithms and compared their accuracy which is 85%,77%, 77.3% respectively. They also used KNN algorithm after the training process to see the reactions of the network which states whether the disease is classified properly or not. Here they compared the precision recall and Fl score support and accuracy of all the models.

2. The main aim of the paper is, as the 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

So, Machine learning and Artificial Intelligence supports in predicting any kind of natural events So in this paper they calculate accuracy of machine learning for predicting heart disease using k-nearest neighbor decision tree, linear regression and SVM by using dataset for training and testing. They also compared the algorithm and their accuracy SVM 83% Decision tree 79% Linear regression 78%, k-nearest neighbor 87%.

3. The system defines that liver diseases are causing a high number of deaths in India and is also considered as a life- threatening disease in the world. As it is difficult to detect liver disease at an early stage So using automated programs using machine learning algorithms we can detect liver disease accurately. They used and compared SVM Decision Tree and Random Forest algorithms and measured precision, accuracy and recall metrics for quantitative measurement. The accuracy is 75% respectively.

# IMPLEMENTATION

**KNN Algorithm**

* The working of the K-NN algorithm is as followed:
* Step-1: Start to select the K value for example k-5 Step-2: Then we will find the Euclidean distance between the points. It is calculated by the as

****

* Step-3: Then we will calculate the Euclidean distance of the nearest neighbor.
* Step-4: Then count the number of the data points in each category For example, find three values for Category B.
* Step-5: Then assign the new point to the category having the maximum number of neighbors. For example Category A has the highest number of neighbors so we will assign the new data point to category A.
* Step-6: So finally, our KNN model is ready.

**Random Forest Algorithm**

* Random Forest working is possible in two phases, first is to create the random forest by merging N decision trees, and second is making predictions for each tree created in the first phase.
* The working of the random forest is as follows:
* Step-1: Firstly, it will select random K data points from the training set.
* Step-2: After selecting k data points then building the decision trees associated with the selected data points (Subsets)
* Step-3: Then choose the number N for decision trees that you want to build
* Step-4: Repeating steps 1 and 2.
* Step-5: Finding the predictions of each decision tree, and assigning the new data points to the category that wins the majority votes.

**Support Vector Machine**

**(SVM)**

* Support vector machine (SVM) algorithm can classify both linear and non-linear data.
* It first maps each data item into an n-dimensional feature space where n is the number of features. It then identifies the hyperplane that separates the data items into two classes while maximising the marginal distance for both classes and minimising the classification errors .
* The marginal distance for a class is the distance between the decision hyperplane and its nearest instance which is a member of that class.
* More formally, each data point is plotted first as a point in an n-dimension space (where n is the number of features) with the value of each feature being the value of a specific coordinate.
* To perform the classification, we then need to find the hyperplane that differentiates the two classes by the maximum margin. Figure provides a simplified illustration of an SVM classifier

.

# figure 2

**Logistic Regression**

* Logistic Regression Logistic Regression is a Machine Learning technique for resolving categorization problems. It's a type of predictive analytic approach based on the probability concept.
* To forecast the probability of a categorical dependent variable, the classification algorithm Logistic Regression is utilized. In logistic regression, the dependent variable is a binary variable with data coded as 1 (yes, True, normal, success, etc.) or 0 (no, False, abnormal, failure, etc).
* The purpose of Logistic Regression is to find a relationship between a set of attributes and the probability of a particular event.
* A Logistic Regression model is similar to a Linear Regression model, except that instead of using a linear function, it uses a more complicated cost function known as the "Sigmoid function."

**Convolutional Neural Network(CNN)**

A Convolutional Neural Network (CNN) is a type of deep learning algorithm that is particularly well-suited for image recognition and processing tasks. It is made up of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers are the key component of a CNN, where filters are applied to the input image to extract features such as edges, textures, and shapes. The output of the convolutional layers is then passed through pooling layers, which are used to down-sample the feature maps, reducing the spatial dimensions while retaining the most important information. The output of the pooling layers is then passed through one or more fully connected layers, which are used to make a prediction or classify the image.

**XGBoost Algorithm**

The working of XGBoost algorithm are as follows:

Step 1: Firstly creating a single leaf tree.

Step 2: Then for the first tree, we have to compute the

average of target variable as prediction and then calculating

the residuals using the desired loss function and then for

subsequent trees the residuals come from prediction that

was there in previous tree.

Step 3: Calculating the similarity score using formula:

where, Hessian is equal to number of residuals; Gradient2 =



squared sum of residuals; λ is a regularization

hyperparameter.

Step 4: Applying similarity score we select the appropriate

node. The higher the similarity score more the homogeneity.

Step 5: Applying similarity score we calculate Information

gain. Information gain help to find the difference between

old similarity and new similarity and tells how much

homogeneity is achieved by splitting the node at a given

point. It is calculated by the formula:



Step 6:Creating the tree of desired length using the above

method pruning and regularization can be done by playing

with the regularization hyperparameter.

Step 7: Then we can predict the residual values using the

Decision Tree you constructed.

Step 8: The new set of residuals is calculated as:

where ρ is the learning rate.



Step 9:Then go back to step 1 and repeat the process for all the trees.

# METHODOLOGY

In multiple disease prediction, it is possible to predict more than one disease at a time. So the user doesn’t need to traverse different sites in order to predict the diseases. To implement multiple disease analyses we are going to use machine learning algorithms and Flask. When the user is accessing this API, the user has to send the parameters of the disease along with the disease name. Flask will invoke the corresponding model and returns the status of the patient.

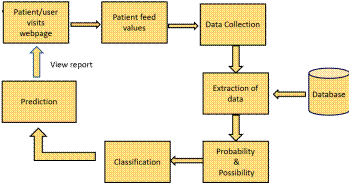
Many of the existing machine learning models for health care analysis are concentrating on one disease per analysis. For example one for heart analysis, one for kidney analysis, one for lung diseases, Diabetes analysis like that. If a user wants to predict more than one disease, he/she has to go through different sites. There is no common system where one analysis can perform more than one disease prediction. Some of the models have lower accuracy which can seriously affect patient’s health. When an organization wants to analyze their patient’s health reports, they have to deploy many models which in turn increases the cost as well as time Some of the existing systems consider very few parameters which can yield false results.

**Functional Requirement**

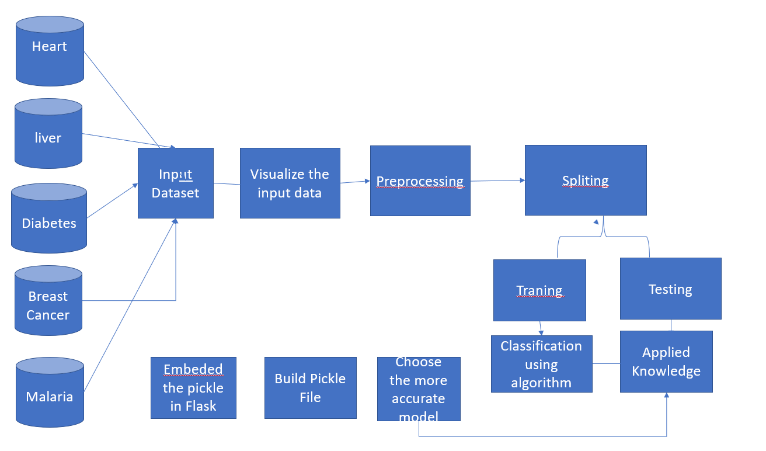
* The system allows the patient to predict the disease
* The user adds the input for the particular disease and based on the trained model of the user input the output will be displayed.

**Non Functional Requirement**

* The website will provide range of the values during the prediction of the disease.
* The website should be reliable and consistent.



ARCHITECTURE DESIGN



ARCHITECTURE

we have experimented on three diseases that is heart, daibetes and liver as these are correlated to each other. The first step is to the dataset for Diabetes, Cancer, Heart, Kidney, Liver, Malaria, Stroke and Pneumonia disease we have imported the dataset respectively. Once we have imported the dataset then visualization of each imputed data takes place. After visualization pre-processing of data takes place where we check for outliers, missing values and also scale the dataset then on the updated dataset we split the data into training and testing .Next is on the training dataset we had applied knn ,xgboost and random forest algorithm and applied knowledge on the classified algorithm using testing dataset. After applying knowledge we will choose the algorithm with the best accuracy for each of the disease .Then we build a pickle file for all the disease and then integrated the pickle file with the Flask framework for the output of the model on the webpage.



Flow Chart

Step 1:

Collecting data for training the ML model is the basic step in the machine learning pipeline. The predictions made by ML systems can only be as good as the data on which they have been trained

Real-world raw data and images are often incomplete, inconsistent and lacking in certain behaviour’s or trends. They are also likely to contain many errors. So, once collected, they are pre-processed into a format the machine learning algorithm can use for the model.

Step 2:

To properly evaluate your machine learning models and select the best one, you need a good validation strategy and solid evaluation metrics picked for your problem.

A good validation (evaluation) strategy is basically how you split your data to estimate future test performance. It could be as simple as a train-test split or a complex stratified k-fold strategy.

Once you know that you can estimate the future model performance, you need to choose a metric that fits your problem. If you understand the classification and regression metrics, then most other complex metrics (in object detection, for example) are relatively easy to grasp.

Step 3:

Training and experimentation on database The first data needed to train machine learning models is known as training data (or a training data set). Machine learning algorithms are taught how to make predictions or perform a task using training assets.

Step 4:

Deployment and analysis Deployment is the process of integrating a machine learning model into an existing production environment in order to make data-driven business decisions. It's one of the last steps in the machine learning process, and it's also one of the most time-consuming.

# **RESULT**

The web application was created successfully we used flask frame work for backend and Html, Bootstrap for frontend. Flask follows the model view controller architecture. The application was tested successfully. The user needs to provide information, which is stored in model of flask, and the data is input for the trained machine learning algorithms. And finally displayed the output in new html page like if the patient has serious issue it displays unhealthy and consult a doctor and if no it says ok your health is great based on the input parameters given by the enduser. The application is tested fine and properly debugged

**VI. CONCLUSION**

The main objective of this project was to create a system

that would predict more than one disease and do so with

high accuracy. Because of this project the user doesn’t need

to traverse different websites which saves time as well.

Diseases if predicted early can increase your life expectancy

**7. FUTURE SCOPE**

1.In the future we can add more diseases in the existing API. 2.We can try to improve the accuracy of prediction in order to decrease the mortality rate

3.Try to make the system user-friendly and provide a

chatbot for normal queries

**8.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)

D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: using analytics

to identify and manage high-risk and high-cost patients,” Health Affairs, vol. 33, no. 7, pp. 1123–1131,

2018.

[3] Implementing WEKA for medical data classification and early disease prediction. “3rd IEEE

International Conference on "Computational Intelligence and Communication Technology" (IEEECICT2017)”.

[4] 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