IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SMART HEALTH PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES

[1]Dr. Nandini C, [2]Antara Mukherjee, [3]Bhoomika M

[1]Vice Principal, Professor & Head, [2]Student, [3]Student

[1]Department of Computer Science and Engineering

[1]Dayananda Sagar Academy of Technology and Management, Bengaluru, Karnataka, India

Abstract: There are various diseases which are considered as deadly and chronic, and it would help the patients if these diseases were diagnosed at the earliest stage. Many complications occur if these diseases remain untreated and unidentified. The aim of developing classifier system using machine learning algorithms is to immensely help to solve the health-related issues by assisting the physicians to predict and diagnose diseases at an early stage. Our proposed system is an end-to-end data science project that is able to predict the chances of getting diseases, based on the blood test report results of the user. In this paper, the risk of getting diseases such as Diabetes, Heart Disease, Kidney Disease, Liver Disease and Breast Cancer, is predicted using various machine learning algorithms. The final output is predicted based on the most accurate machine learning algorithm. A web application is designed where the user will have the convenience to select the diseases they want consultancy on and enter their blood test report details. Then, the system uses the most accurate model which can prognosticate the likelihood of the selected disease in patients.

Index Terms **- Machine Learning, Health Care, Classification, Support Vector Machine, Accuracy**

I. INTRODUCTION

Machine Learning is a very promising approach which helps in early diagnosis of disease and might help the practitioners in decision making for diagnosis. We explore the landscape of recent advances to address the challenges model interpretability in healthcare and also describe how one would go about choosing the right interpretable machine learning algorithm for a given problem in healthcare. Machine learning and artificial intelligence hold the potential to transform healthcare and open up a world of incredible promise.

This project extensively covers the definitions, nuances, challenges, and requirements for the design of interpretable and explainable machine learning models and systems in healthcare. We discuss many uses in which interpretable machine learning models are needed in healthcare and how they should be deployed. Prediction by a traditional bio-medical machine learning model typically involves some supervised algorithm which uses guidance data with the label for the prediction of the models. Classification strategies are broadly used in the medical field for classifying data into different classes according to some constrains comparatively an individual classifier.

Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. Diabetes is not only affected by various factors like height, weight, hereditary factor and insulin but the major reason considered is sugar concentration among all factors. The early identification is the only remedy to stay away from the complications. Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48 [1], Support Vector Machine, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithms [2] works better in diagnosing different diseases. Data Mining [2], and Machine learning algorithms gain its strength due to the capability of managing a large amount of data to combine data from several different sources and integrating the background information in the study. This research work focuses on pregnant women suffering from diabetes. A diabetes prediction model is proposed for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing Pima Indians Diabetes Database (PIDD) dataset [3]. Further, a pipeline model for diabetes prediction intended towards improving the accuracy of classification is imposed.

Breast cancer has now overtaken lung cancer as the most commonly diagnosed cancer in women worldwide, according to statistics released by the International Agency for Research on Cancer (IARC) in December 2020. In the past two decades, the overall number of people diagnosed with cancer nearly doubled, from an estimated 10 million in 2000 to 19.3 million in 2020. This paper mainly gives a comparison between the performance of five classifiers: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbors which according to research community are among the most influential data mining algorithms and among the top 10 data mining algorithms, on the Breast Cancer Wisconsin Diagnostic dataset. After obtaining the results, a performance evaluation and comparison is carried out between these different algorithms. Our objective is to

predict and diagnosis breast cancer, using machine-learning algorithms, and find out the most effective based on the performance of each classifier in terms of confusion matrix, accuracy, precision and sensitivity [4].

Heart disease is the leading cause of death in the world over the past 10 years (World Health Organization 2007). Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. However, using data mining technique can reduce the number of tests that are required. In order to reduce number of deaths from heart diseases there have to be a quick and efficient detection technique. Decision Tree is one of the effective data mining methods used. This research compares different algorithms of Decision Tree classification seeking better performance in heart disease diagnosis using WEKA. The algorithms which are tested are J48 algorithm, Logistic model tree algorithm and Random Forest algorithm. The existing datasets of heart disease patients from Cleveland database of UCI repository is used to test and justify the performance of decision tree algorithms [1]. This dataset consists of 303 instances and 76 attributes. Subsequently, the classification algorithm that has optimal potential will be suggested for use in sizeable data. The goal of this study is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where this presence is valued from no presence to likely presence.

Diagnosis of liver disease at a preliminary stage is important for better treatment. It is a very challenging task for medical researchers to predict the disease in the early stages owing to subtle symptoms. Often the symptoms become apparent when it is too late. To overcome this issue, this project aims to improve liver disease diagnosis using machine learning approaches. The main objective of this research is to use classification algorithms to identify the liver patients from healthy individuals [5]. The dataset used is The Indian Liver Patient Dataset (ILPD) which was selected from UCI Machine learning repository for this study. It is a sample of the entire Indian population collected from Andhra Pradesh region and comprises of 585 patient data [6].

The disability of the kidneys to perform their regular blood filtering function and others is called chronic kidney disease (CKD). The term “chronic” describes the slow degradation of the kidney cells over a long period of time. This disease is a major kidney failure where the kidney sans blood filtering process and there is a heavy fluid build-up in the body. This leads to alarming increase of potassium and calcium salts in the body. Existence of high levels of these salts result in various other ailments in the body [7]. But machine learning can be our hope in this problem as it is best in prediction and analysis. The dataset used is the chronic kidney disease dataset where the data was taken over a 2-month period in India with 25 features (e.g., red blood cell count, white blood cell count, etc) consisting of 400 rows. We are going to use various machine learning techniques like Decision Tree and SVM to build a model with maximum accuracy of predicting whether chronic kidney disease is present or not [8].

The modern approach to healthcare is to prevent the disease with early intervention rather than go for treatment after diagnosis. Traditionally, physicians or doctors use a risk calculator to assess the possibility of disease development. The motive of this study is to design models which can prognosticate the likelihood of various diseases in patients with maximum accuracy. We plan to create an end user support and online consultation system. This paper shows us a way of building an application which can help to solve the health-related issues by assisting the physicians and patients to predict and diagnose diseases at an early stage.

II. LITERATURE SURVEY

The analysis of related work gives results on various healthcare datasets, where analysis and predictions were carried out using various methods and techniques. Various prediction models have been developed and implemented by various researchers using variants of data mining techniques, machine learning algorithms or also combination of these techniques.

Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel [1] in their research compare different algorithms of Decision Tree classification seeking better performance in heart disease diagnosis using WEKA. The goal of this study is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where this presence is valued from no presence to likely presence. It is observed that applying reduced error pruning to J48 results in higher performance while without pruning, it results in lower performance. The best algorithm J48 based on UCI data has the highest accuracy i.e., 56.76% and the total time to build model is 0.04 seconds. But only a marginal success is achieved in the creation of predictive model for heart disease patients. Hence, there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease. Further work involves development of system using the mentioned methodology to be use for checking the imbalance with other data mining models, to explore different rules for better efficiency and ease of simplicity and to make use of Multivariate Decision Tree approach on smaller and larger amount of data.

Deepti Sisodia, Dilip Singh Sisodia [2] designed a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. They used three machine learning classification algorithms: Decision Tree, SVM and Naive Bayes to detect diabetes at an early stage. Experiments are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. But, accuracy of 76% is achieved in the creation of predictive model for diabetes disease, hence they mentioned that there is a need for more complex models to increase the accuracy of predicting these models. For future work, they stated that the work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

Aishwarya Mujumdar, Dr. Vaidehi V [3] studied various machine learning algorithms and applied them on the diabetes dataset and the classification has been done using various algorithms of which Logistic Regression gives highest accuracy of 96%. Application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8%. We have seen comparison of machine learning algorithm accuracies with two different datasets. It is clear that the model improves accuracy and precision of diabetes prediction with this dataset compared to existing dataset. Further this work can be extended to find how likely non-diabetic people can have diabetes in next few years.

Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar, Rachida Ait Abdelouhahid, Olivier Debauche [4] used five machine learning algorithms are applied: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbors (KNN) on the Breast Cancer Wisconsin Diagnostic dataset. The main objective of their research paper was to predict and diagnosis breast cancer, using machine-learning algorithms. After an accurate comparison between the models, it was found that Support Vector Machine achieved a higher efficiency of 97.2% and outperforms all other algorithms. But a weakness was found that all the results obtained are related just to the WBCD database and there is no evidence if it works on other datasets. For future work, they suggested to apply the already applied machine learning algorithms and other machine learning algorithms using new parameters on larger data sets with more disease classes to obtain higher accuracy.

Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac [5] have proposed methods for diagnosing liver disease in patients using machine learning techniques. The four machine learning techniques that were used include SVM, Logistic Regression, KNN and Artificial Neural Network. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. ANN was the model that resulted in the highest accuracy with an accuracy of 98%. Comparing this work with the previous research works, it was discovered that ANN proved highly efficient. A GUI, which can be used as a medical tool by hospitals and medical staff was implemented using ANN.

Rakshith D B, Mrigank Srivastava, Ashwani Kumar, Gururaj S P [6] for their research paper had the main aim to predict liver disease using different classification algorithms. The algorithms used for this purpose of work are Logistic Regression, K-Nearest Neighbour and Support Vector Machines. The system predicts the results with 90% accuracy for the dataset Indian Liver Patient Dataset (ILPD). The dataset used has approximately 600 rows, which does not prepare the model for predicting various abnormal cases, hence this model needs to be utilized for much larger datasets with more attributes to cause the model to perform even more accurately. In future, collection of recent data from various regions across the world for liver disease diagnosis can be applied and these models can be incorporated these into an Android app or iOS app.

S. Revathy, B. Bharathi, P. Jeyanthi, M. Ramesh [7] used knowledge discovery as an important application of data mining which involves various stages of processing. The paper tries to propose a datamining framework for knowledge discovery on the CKD datasets. Data preparation and pre-processing is done using the traditional methods of data mining process. Three machine learning algorithms: Decision tree, Random Forest and Support Vector Machines are used to predict the early occurrence of CKD. High accuracy can be achieved using Random Forest algorithm (99.16%). For future work, the comparison can be done based on the time of execution, feature set selection as the improvisation of this research. It can also be incorporated into a website, and these app and website will be highly beneficial for a large section of society.

Siddheshwar Tekale, Pranjal Shingavi, Sukanya Wandhekar, Ankit Chatorikar [8] have studied different machine learning algorithms and analysed 14 different attributes related to CKD patients. From the results analysis, it is observed that the decision tree algorithms give the accuracy of 91.75% and SVM gives accuracy of 96.75%. The advantage of this system is that, the prediction process is less time consuming. It will help the doctors to start the treatments early for the CKD patients and also it will help to diagnose more patients within a less time period. Limitations of this study are the strength of the data is not higher because of the size of the data set and the missing attribute values. To build a machine learning model targeting chronic kidney disease with overall accuracy of 99.99%, will need millions of records with zero missing values.

III. METHODOLOGY

A machine learning model is built by learning and generalizing from training data, then applying that acquired knowledge to new data it has never seen before to make predictions and fulfill its purpose. The process of gathering data depends on the type of project we desire to make, if we want to make a machine learning project that uses real-time data, then we can build an IoT system that using different sensors data. The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem data pre-processing is done. Data pre-processing is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis. For training a model we initially split the model into training data and testing data. You train the classifier using training data set, then test the performance of your classifier on unseen test data set. The selection of the model type is our next course of action once we are done with the data-centric steps. There are various existing models developed by data scientists which can be used for different purposes. These models are designed with different goals in mind. For instance, some models are more suited to dealing with texts while another model may be better equipped to handle images. The last step is prediction, which refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome.

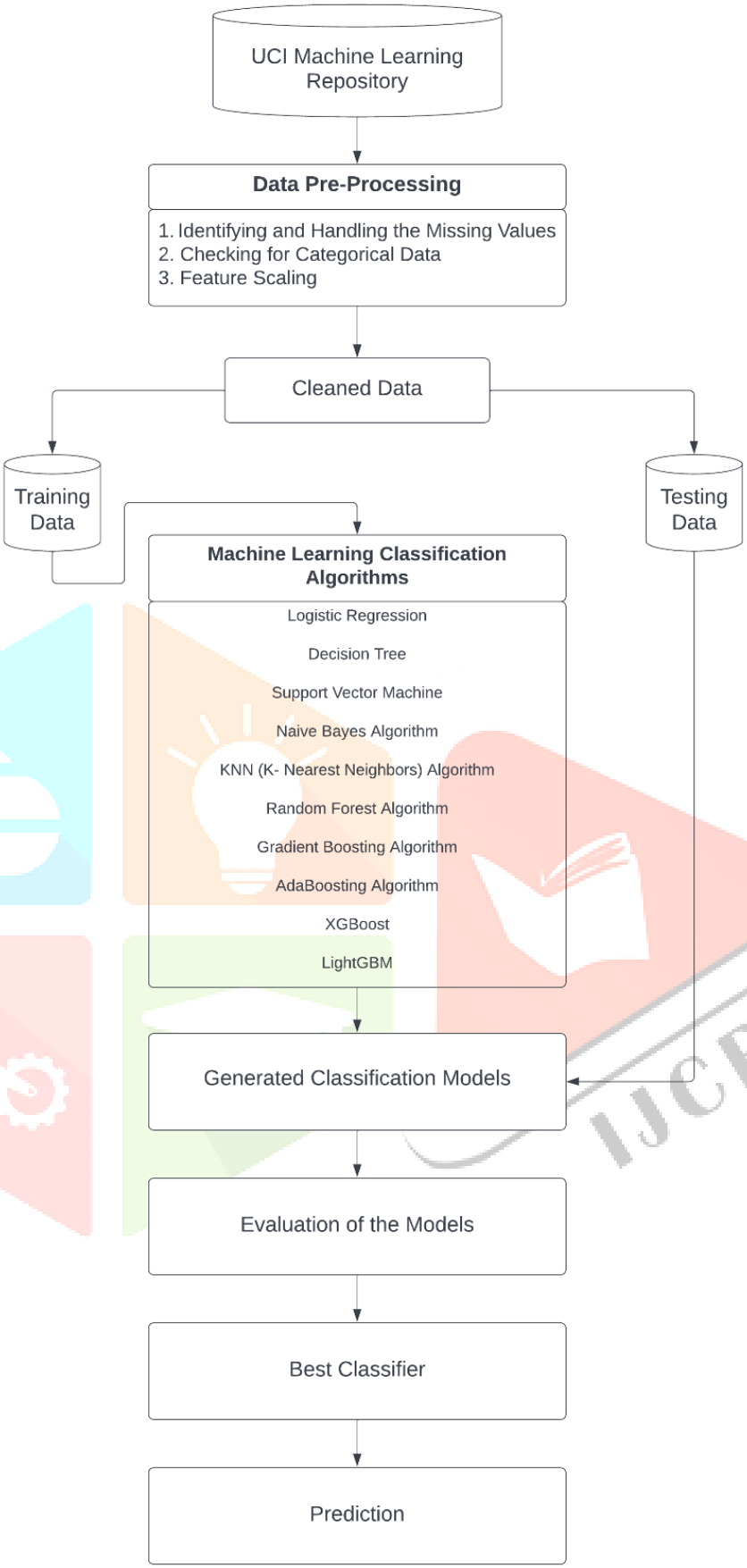

Fig -1: Proposed Model Diagram

3.1 Data Acquisition

A dataset is a collection of data in which data is arranged in some order. To build and develop Machine Learning models, you must first acquire the relevant dataset. This dataset will be comprised of data gathered from multiple and disparate sources which are then combined in a proper format to form a dataset. UCI Machine learning repository is one of the great sources of machine learning datasets. This repository contains databases, domain theories, and data generators that are widely used by the machine learning community for the analysis of ML algorithms. Since the year 1987, it has been widely used by students, professors, researchers as a primary source of machine learning dataset.

3.1.1 Diabetes Dataset Description

Pima Indians Diabetes Database (PIDD) dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Table -1: Pima Indians Diabetes Database (PIDD) Dataset

No.	Attributes	Type	Description
1	Pregnancies	Continuous	Number of times pregnant
2	Glucose	Continuous	Plasma glucose concentration
3	BloodPressure	Continuous	Diastolic blood pressure (mm Hg)
4	SkinThickness	Continuous	Triceps skin fold thickness (mm)
5	Insulin	Continuous	2-Hour serum insulin (mu U/ml)
6	BMI	Continuous	Body mass index (weight in kg/ (height in m) ^2)
7	DiabetesPedigreeFunction	Continuous	Diabetes pedigree function
8	Age	Continuous	Age (years)
9	Outcome	Discrete	Class variable (0=No Presence, 1=More likely have diabetes)

All features are numeric values. The last column, Outcome, is the label (with '1' representing presence of disease and '0' representing absence of disease). Total number of data points is 2000, with 1316 diabetes patient records and 684 non-diabetes patient records. Variables such as BloodPressure, BMI, Glucose, Insulin, SkinThickness logically cannot be zero. This means that missing values in the dataset are filled with 0 instead of NaN. Then, actual null values appear in the dataset. These values can be filled with mean values of the column. Then we scale the dataset to normalize all values.

3.1.2 Heart Disease Dataset Description

Cleveland dataset from UCI repository is used. The dataset has 76 attributes and 303 records. However, only 13 attributes are used for this study and testing.

Table -2: Cleveland Heart Disease Dataset

No.	Attributes	Type	Description
1	age	Continuous	Age in years
2	restecg	Discrete	Resting electrocardiographic results (0,1,2)
3	sex	Discrete	Sex of the patient (0=Female, 1=Male)
4	cp	Discrete	Chest pain type (1=Typical angina, 2=Atypical angina, 3=Non-anginal pain, 4=Asymptom)
5	trestbps	Continuous	Resting blood pressure (in mm Hg)
6	chol	Continuous	Serum cholesterol in mg/dl
7	fbs	Discrete	Fasting blood sugar>120 mg/dl (1=True, 0=False)
8	exang	Discrete	Exercise induced angina (1=Yes, 0=No)
9	thalach	Continuous	Maximum heart rate achieved
10	oldpeak	Continuous	Depression induced by exercise relative to rest
11	slope	Discrete	The slope of the peak exercise segment (1=Up sloping, 2=Flat, 3=Down sloping)
12	ca	Continuous	Number of major vessels colored by fluoroscopy that ranged between 0 and 3
13	thal	Discrete	A blood disorder called thalassemia (3=Normal, 6=Fixed defect, 7=Reversible defect)
14	target	Discrete	Diagnosis classes (0=No Presence, 1=More likely have heart disease)

All features are numeric values. The last column, target, is the label (with '1' representing presence of disease and '0' representing absence of disease). Total number of data points is 303, with 138 heart disease patient records and 165 non-heart disease patient records. We scale the dataset to normalize all values.

3.1.3 Breast Cancer Dataset Description

Breast Cancer Wisconsin Diagnostic dataset from University of Wisconsin Hospitals Madison Breast Cancer Database. The features of dataset are computed from a digitized image of a breast cancer sample obtained from fine-needle aspirate (FNA). The characteristics of the cell nuclei present in the image are determined from these features.

Table -3: Wisconsin Breast Cancer (Diagnostics) Dataset

No.	Attributes	Type	Description
1	radius	Continuous	Mean of distances from center to points on the perimeter
2	texture	Continuous	Standard deviation of gray-scale values
3	perimeter	Continuous	Perimeter
4	area	Continuous	Area
5	smoothness	Continuous	Local variation in radius lengths
6	compactness	Continuous	Perimeter^2 / area - 1.0
7	concavity	Continuous	Severity of concave portions of the contour
8	concave points	Continuous	Number of concave portions of the contour
9	symmetry	Continuous	Symmetry
10	fractal dimension	Continuous	"Coastline approximation" - 1
11	diagnosis	Discrete	Diagnosis of breast tissues (M=Malignant, B=Benign)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. All features are numeric values. Total number of data points is 569, with 357 benign and 212 malignant records. We scale the dataset to normalize all values.

3.1.4 Indian Liver Patient Dataset Description

Indian Liver Patient Dataset is originally collected from North East of Andhra Pradesh, India. This data set contains 441 male patient records and 142 female patient records.

Table -4: Indian Liver Patient Dataset

No.	Attributes	Type	Description
1	Age	Continuous	Age of the patient
2	Gender	Discrete	Gender of the patient (Female, Male)
3	Total_Bilirubin	Continuous	Total Bilirubin
4	Direct_Bilirubin	Continuous	Direct Bilirubin
5	Alkaline_Phosphotase	Continuous	Alkaline Phosphotase
6	Alamine_Aminotransferase	Continuous	Alamine Aminotransferase
7	Aspartate_Aminotransferase	Continuous	Aspartate Aminotransferase
8	Total_Protiens	Continuous	Total Protiens
9	Albumin	Continuous	Albumin
10	Albumin_and_Globulin_Ratio	Continuous	Albumin and Globulin Ratio
11	Dataset	Discrete	Diagnosis classes (1=Presence of disease, 2=Absence of disease).

All features, except Gender are numeric values. The last column, Dataset, is the label (with ’1’ representing presence of disease and ’2’ representing absence of disease). Total number of data points is 583, with 416 liver disease patient records and 167 non-liver disease patient records. In the description of this dataset, it is observed that some values are Null for the Albumin and Globulin Ratio column. The columns which contain null values are replaced with mean values of the column.

3.1.5 Chronic Kidney Disease Dataset Description

Chronic Kidney Disease dataset consists of data the was taken over a 2-month period in India with 25 features (e.g., red blood cell count, white blood cell count, etc.) consisting of 400 rows.

Table -5: Chronic Kidney Disease Dataset

No.	Attributes	Type	Description
1	age	Continuous	Age
2	bp	Continuous	Blood pressure
3	sg	Continuous	Specific gravity
4	al	Continuous	Albumin
5	su	Continuous	Sugar
6	rbc	Continuous	Red blood cells
7	pc	Continuous	Pus cell
8	pcc	Continuous	Pus cell clumps
9	ba	Continuous	Bacteria
10	bgr	Continuous	Blood glucose random
11	bu	Continuous	Blood urea
12	sc	Continuous	Serum creatinine
13	sod	Continuous	Sodium
14	pot	Continuous	Potassium
15	hemo	Continuous	Haemoglobin
16	pcv	Continuous	Packed cell volume
17	wc	Continuous	White blood cell count
18	rc	Continuous	Red blood cell count
19	htn	Continuous	Hypertension
20	dm	Continuous	Diabetes mellitus
21	cad	Continuous	Coronary artery disease
22	appet	Continuous	Appetite
23	pe	Continuous	Pedal edema
24	ane	Continuous	Anemia
25	classification	Discrete	Class (ckd, notckd)

All features, except classification, are numeric values. The last column, classification, is the label (with 'ckd' representing presence of disease and 'notckd' representing absence of disease). Total number of data points is 400, with 250 kidney disease patient records and 150 non-kidney disease patient records. Then we scale the dataset to normalize all values.

3.2 Data Pre-Processing

Data pre-processing in Machine Learning is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data. It refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. In simple words, data pre-processing in Machine Learning is a data mining technique that transforms raw data into an understandable and readable format.

3.2.1 Identifying and Handling the Missing Values

In data preprocessing, it is pivotal to identify and correctly handle the missing values, failing to do this, you might draw inaccurate and faulty conclusions and inferences from the data. In real life, many datasets will have many missing values, so dealing with them is an important step. For filling missing values, there are many methods available. For choosing the best method, you need to understand the type of missing value and its significance, before you start filling/deleting the data.

3.2.2 Checking for Categorical Data

Data in the dataset has to be in a numerical form so as to perform computation on it. Since Machine learning models contain complex mathematical computation, we can't feed them a non-numerical value. So, it is important to convert all the text values into numerical values.

3.2.3 Feature Scaling

The values of the raw data vary extremely and it may result in biased training of the model or may end up increasing the computational cost. So, it is important to normalize them. Feature scaling is a technique that is used to bring the data value in a shorter range. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no any variable dominates the other variable.

3.3 Build Model

For training a model we initially split the model into two sections: training data and testing data. You train the classifier using training data set, then test the performance of your classifier on unseen test data set. An important point to note is that during training the classifier only the training set is available. The test data set must not be used during training the classifier. The test set will only be available during testing the classifier. The selection of the model type is our next course of action once we are done with the data-centric steps. There are various existing models developed by data scientists which can be used for different purposes. These models are designed with different goals in mind. For instance, some models are more suited to dealing with texts while another model may be better equipped to handle images.

3.3.1 Training and Testing Data

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. To detect a machine learning model behavior, we need to use observations that aren't used in the training process. Otherwise, the evaluation of the model would be biased. Using the training observations for model evaluation is like giving a class of students a set of questions, then asking some of the questions in the final exam. We can't know whether the pupils really learned the subject or just memorized some specific data. The simplest method is to divide the whole dataset into two sets. Then use one for training and the other for model evaluation. Generally, the training and test data set is split into an 80:20 ratio. Thus, 20% of the data is set aside for testing purposes. We train the machine learning models using these observations. In other words, we feed these observations into the model to update its parameters during the learning phase. We test the machine learning model after the training phase is complete, using the observations from the test set. This way, we measure how the model reacts to new observations.

3.3.2 Machine Learning Classification Algorithms

The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into. The classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.

3.3.2.1 Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). In Logistic regression, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

3.3.2.2 Decision Tree

Decision Tree algorithm in machine learning is one of the most popular algorithms in use today; this is a supervised learning algorithm that is used for classifying problems. It works well classifying for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

3.3.2.3 Support Vector Machine

SVM algorithm is a method of classification algorithm in which you plot raw data as points in an n-dimensional space (where n is the number of features you have). The value of each feature is then tied to a particular coordinate, making it easy to classify the data.

3.3.2.4 Naive Bayes

A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features are related to each other, a Naive Bayes classifier would consider all of these properties independently when calculating the probability of a particular outcome. A Naive Bayesian model is easy to build and useful for massive datasets. It's simple and is known to outperform even highly sophisticated classification methods.

3.3.2.5 K- Nearest Neighbors

KNN algorithm can be applied to both classification and regression problems. It's a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. The case is then assigned to the class with which it has the most in common. A distance function performs this measurement.

3.3.2.6 Random Forest Classifier

A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

3.3.2.7 Gradient Boosting Algorithm

Boosting algorithms are used when massive loads of data have to be handled to make predictions with high accuracy. Boosting is an ensemble learning algorithm that combines the predictive power of several base estimators to improve robustness. It is a technique of producing an additive predictive model by combining various weak predictors, typically Decision Trees.

3.3.2.8 AdaBoosting Algorithm

The AdaBoost algorithm involves using very short (one-level) decision trees as weak learners that are added sequentially to the ensemble. Each subsequent model attempts to correct the predictions made by the model before it in the sequence. It combines multiple weak or average predictors to build a strong predictor.

3.3.2.9 XGBoost

It is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

3.3.2.10 LightGBM

It is a fast, distributed, high performance gradient boosting framework based on decision tree algorithms, used for ranking, classification and many other machine learning tasks.

3.4 Evaluation of the Models

Evaluation metrics are used to measure the quality of the model. One of the most important topics in machine learning is how to evaluate your model. When you build your model, it is very crucial to measure how accurately it predicts your expected outcome. We have different evaluation metrics for a different set of machine learning algorithms. For evaluating classification models, we use classification metrics such as Accuracy, Precision, Sensitivity and Specificity. Evaluation metrics can help you assess your model's performance, monitor your ML system in production, and control your model to fit your business needs. Our goal is to create and select a model which gives high accuracy on out-of-sample data. It's very crucial to use multiple evaluation metrics to evaluate your model because a model may perform well using one measurement from one evaluation metric while may perform poorly using another measurement from another evaluation metric.

3.4.1 Confusion Matrix

A confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.

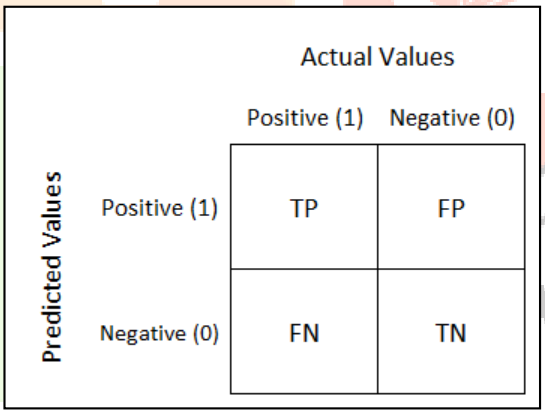

Fig -2: Confusion Matrix

3.4.2 Accuracy

The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

$$\text{Accuracy} = (\text{No. of TP} + \text{No. of TN}) / (\text{No. of TP} + \text{No. of TN} + \text{No. of FP} + \text{No. of FN}) \quad (1)$$

3.4.3 Sensitivity

Sensitivity is also referred as True positive rate i.e., the proportion of positive tuples that are correctly identified.

$$\text{Sensitivity} = \text{No. of TP} / (\text{No. of TP} + \text{No. of FN}) \quad (2)$$

3.4.4 Precision

Precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives).

$$\text{Precision} = \text{No. of TP} / (\text{No. of TP} + \text{No. of FP}) \quad (3)$$

3.4.5 Specificity

Specificity is the True negative rate that is the proportion of negative tuples that are correctly identified.

$$\text{Specificity} = \text{No. of TN} / (\text{No. of TN} + \text{No. of FP}) \quad (4)$$

Where, TP: True Positive, FP: False Positive, FN: False Negative and TN: True Negative

3.5 Prediction

Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome. The algorithm will generate probable values for an unknown variable for each record in the new data, allowing the model builder to identify what that value will most likely be.

IV. RESULTS

Training accuracy is usually the accuracy you get if you apply the model on the training data, while testing accuracy is the accuracy for the testing data. It's useful to compare these to identify overtraining. The results of each of the classification algorithms that has been generated is evaluated using accuracy of both the training and testing datasets for each of the disease datasets.

4.1 Diabetes

Table -6: Results of Classification Algorithms for Diabetes Dataset

No.	Algorithms	Training accuracy	Test accuracy
1	Random Forest	100%	98.50%
2	LightGBM	100%	98.25%
3	Decision Tree	100%	97.50%
4	Gradient Boosting	92.50%	89.75%
5	XGBoost	90.93%	88.25%

As clearly summarized in the table, Random Forest Classifier gave the best results.

4.2 Heart Disease

Table -7: Results of Classification Algorithms for Heart Disease Dataset

No.	Algorithms	Training accuracy	Test accuracy
1	Support Vector Machine	91.32%	83.60%
2	Gradient Boosting	99.58%	81.96%
3	Naive Bayes	84.29%	81.96%
4	Random Forest	100%	80.32%
5	XGBoost	92.56%	80.32%

As clearly summarized in the table, Support Vector Machine gave the best results.

4.3 Breast Cancer

Table -8: Results of Classification Algorithms for Breast Cancer Dataset

No.	Algorithms	Training accuracy	Test accuracy
1	Support Vector Machine	98.24%	98.24%
2	K-Nearest Neighbors	97.80%	98.24%
3	AdaBoost	100%	96.49%
4	Logistic Regression	98.68%	95.61%
5	LightGBM	100%	95.61%

As clearly summarized in the table, Support Vector Machine gave the best results.

4.4 Liver Disease

Table -9: Results of Classification Algorithms for Liver Disease Dataset

No.	Algorithms	Training accuracy	Test accuracy
1	Support Vector Machine	71.45%	70.94%
2	Logistic Regression	72.10%	69.23%
3	XGBoost	83.26%	63.24%
4	Random Forest	100%	62.39%
5	AdaBoost	83.69%	60.68%

As clearly summarized in the table, Support Vector Machine gave the best results.

4.5 Kidney Disease

Table -10: Results of Classification Algorithms for Kidney Disease Dataset

No.	Algorithms	Training accuracy	Test accuracy
1	LightGBM	100%	100%
2	Support Vector Machine	100%	100%
3	AdaBoost	100%	99.37%
4	Random Forest	100%	98.75%
5	Gradient Boosting	100%	98.75%

As clearly summarized in the table, LightGBM gave the best results.

V. CONCLUSION

A novel processing method has been proposed in this study to effectively predict various diseases incidence. In this project, we have proposed methods for diagnosing diabetes, breast cancer, heart disease, liver disease and kidney disease in patients using machine learning techniques. The ten machine learning techniques that were used include SVM, Logistic Regression, KNN, Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, AdaBoosting, XGBoost and LightGBM. The system was implemented using all the models and their performance was evaluated. For further work, we need to perform Receiver-Operating Characteristic (ROC curve) analysis for evaluating diagnostic tests and predictive models, calculate the computational complexity of all the algorithms and find out which is the most efficient, create a complete user interface for users to easily navigate and predict disease based on the models we have built, try to improve some of the models further by hyperparameter tuning and perform Exploratory Data Analysis to find the relationships among features and target.

REFERENCES

[1] Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel, 2016. Heart Disease Prediction Using Machine Learning and Data Mining Technique. Department of Computer Science and Engineering, Nirma University, Gujarat, India. DOI: 10.090592/IJCSC.2016.018.

[2] Deepti Sisodia, Dilip Singh Sisodia, 2018. Prediction of Diabetes using Classification Algorithms. International Conference on Computational Intelligence and Data Science. ICCIDS 2018.

[3] Aishwarya Mujumdar, Dr. Vaidehi V, 2019. Diabetes Prediction using Machine Learning Algorithms. Vellore Institute of Technology, Chennai, India, Mother Teresa Women's University, Kodaikanal, India. ICRTAC 2019.

[4] Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar, Rachida Ait Abdelouhahid, Olivier Debauche, 2021. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. International Workshop on Edge IA-IoT for Smart Agriculture (SA2IOT) August 9-12, 2021, Leuven, Belgium.

[5] Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac, 2018. Diagnosis of Liver Disease Using Machine Learning Techniques. Dept. of Computer Science and Engineering, MACE, Kerala, India. IRJET 2018. p-ISSN: 2395-0072.

[6] Rakshith D B, Mrigank Srivastava, Ashwani Kumar, Gururaj S P, 2021. Liver Disease Prediction System using Machine Learning Techniques. Department of Computer Science and Engineering, Siddaganga Institute of Technology, Tumkur, India. ISSN: 2278-0181, IJERTV10IS060460.

[7] S. Revathy, B. Bharathi, P. Jeyanthi, M. Ramesh, 2019. Chronic Kidney Disease Prediction using Machine Learning Models. Blue Eyes Intelligence Engineering & Sciences Publication. ISSN: 2249-8958, Volume-9 Issue-1, October 2019. DOI: 10.35940/ijeat.A2213.109119.

[8] Siddheshwar Tekale, Pranjal Shingavi, Sukanya Wandhekar, Ankit Chatorikar, 2018. Prediction of Chronic Kidney Disease Using Machine Learning Algorithm. Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati. DOI 10.17148/IJARCCE.2018.71021.