iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5373

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Predictive Analysis on Diabetes, Liver and Kidney Diseases using Machine Learning

Shruti Katiyar[1], Shruti Jain[2]
[1, 2]*Student, Computer Science, Raj Kumar Goel Institute of Technology, Uttar Pradesh, India*

Abstract: ***Healthcare data analysis is becoming one of the most promising research areas. Diabetes, Liver Disease and Chronic Kidney Disease combinedly affects a mass of world's population. The objective of this briefing is to develop an efficient decision support system to predict the possibility of a disease using the techniques of Machine Learning. Machine Learning is used to discover patterns in the data, detect and analyze trends and then make predictions with the help of algorithms. It provides methods, techniques and tools that can help in solving diagnostic problems in a variety of medical domains e.g. prediction of disease progression, extraction of medical knowledge for outcome research, therapy planning and support, and for the overall patient management. It offers a principled approach for developing sophisticated, automatic, and objective algorithms for biomedical data. This paper mainly focuses on diagnostically predict the possibility of mainly three diseases: Diabetes, Liver Disease and Chronic Kidney Disease.***

Keywords: Machine Learning (ML), Naïve Bayes Classification, Kernel Support Vector Classification, Decision Tree Classifier, Random Forest Classification, Logistic Regression Model, XGBoost Classifier, Diabetes, Liver Disease, Chronic Kidney Disease.

I. INTRODUCTION

Diabetes, Liver Disease and Chronic Kidney Disease are affecting millions of people worldwide. But with early diagnosis and treatment, it's possible to slow or stop the progression of these diseases.

Diabetes is a serious, long-term condition with a major impact on the lives and well-being of individuals, families, and societies worldwide. The global diabetes prevalence in 2019 is estimated to be 9.3% (463 million people), rising to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045. One in two (50.1%) people living with diabetes do not know that they have diabetes Just under half a billion people are living with diabetes worldwide and the number is projected to increase by 25% in 2030 and 51% in 2045. [1]

Liver Disease has emerged as a growing public health problem worldwide. Over the past several decades, it has relentlessly risen to become one of the leading causes of death and illness. Liver disease accounts for approximately 2 million deaths per year worldwide. Global prevalence of liver disease from autopsy studies ranges from 4.5% to 9.5% of the general population. Hence, it can be said that more than fifty million people in the world, taking the adult population, would be affected with chronic liver disease. [2]

Chronic kidney diseases (CKDs) are the most common forms of kidney disease all around the world. The incidence of CKD is rising is mainly driven by population aging as well as by a global rise in hypertension, metabolic syndrome, and metabolic risk factors. Chronic kidney disease is a worldwide health crisis. 10% of the population worldwide is affected by chronic kidney disease (CKD), and millions die each year because they do not have access to affordable treatment. [3]

The objective of this study is to assess the efficiency and accuracy of the used ML models for the prediction of these three diseases. This study focuses on structured data and for the best possible output classification and prediction, advanced machine learning algorithms like Random Forest, Decision Tree, Naive Bayes Classification, XGBoost Classifier, Kernel Support Vector Classification and Logistic Regression has been used.

II. MACHINE LEARNING

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. It focuses mainly on machine learning from their experience and making predictions based on its experience .A formal definition of machine learning is given by Mitchel: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. [4]

Machine learning is sub-categorized to three types:

A. *Supervised Learning*

In supervised learning, the system must "learn" inductively a function called target function, which is an expression of a model describing the data. [5] A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data
In this paper, prediction using supervised learning models has been done.

B. *Unsupervised Learning*

In unsupervised learning, the system tries to discover the hidden structure of data or associations between variables. In that case, training data consists of instances without any corresponding labels. [5]

C. *Reinforcement Learning*

In Reinforcement Learning, system attempts to learn through direct interaction with the environment so as to maximize some notion of cumulative reward. [5]

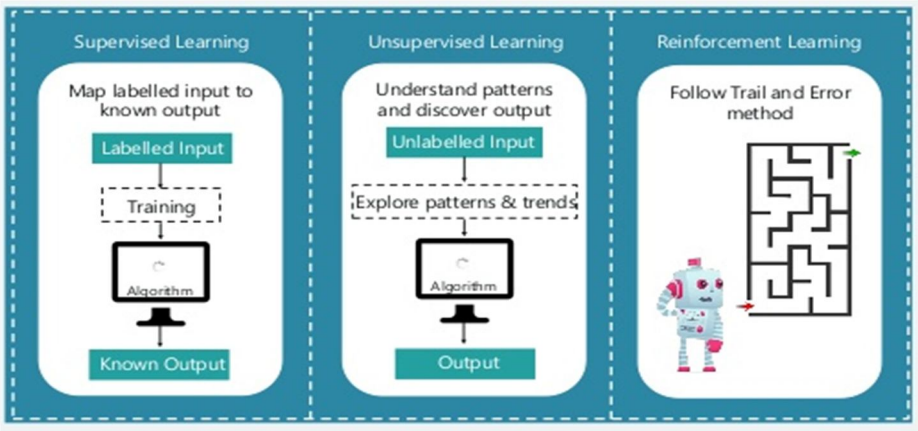


Fig.1 Approaches of different types of Machine Learning

III. ALGORITHMS

To find hidden insights without needing explicit programming, Machine learning uses algorithms which learn from previous data to help produce reliable and repeatable decisions. Machine Learning at its core is just is a collection of algorithms. ML Algorithms leverage knowledge of statistics, probability, calculus, vector algebra, matrices, optimization techniques etc. Six classification algorithms has been used it this system which are as follows:

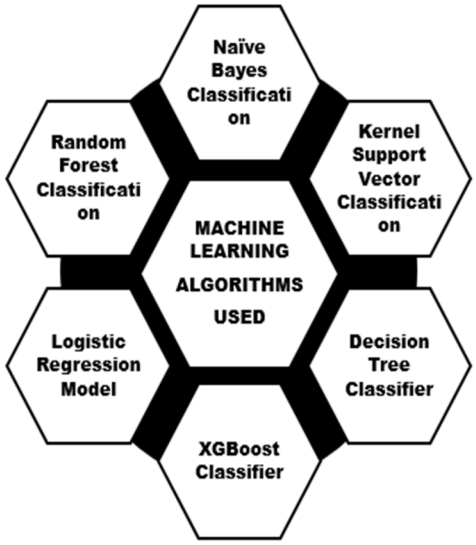
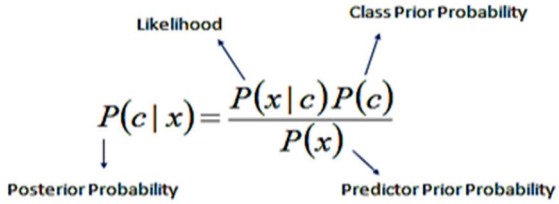


Fig. 2 Different algorithms used in this project

A. Naïve Bayes Classification

Naïve Bayes mainly targets the text classification industry. It is mainly used for clustering and classification purpose [6]. The underlying architecture of Naïve Bayes depends on the conditional probability. It creates trees based on their probability of happening. The Naïve Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. This model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Bayes theorem provides a way of calculating the posterior probability P(C/X) of class from P(C) is the prior probability of class.



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Fig. 3 Equation for Naïve Bayes

P(X) is the prior probability of predictor and P(X/C) is the likelihood which is the probability of predictor given class. Naïve Bayes classifier assumes that the effect of the value of a predictor (X) on a given class(C) is independent of the values of other predictors called conditional independence. [7]

B. Kernel Support Vector Classification

SVM are powerful yet flexible supervised machine learning algorithms. Generally, it is considered to be a classification approach, but it can be used both for classification and regression process. SVM is able to handle multiple continuous and categorical variables. SVM constructs a decision boundary or hyperplane that divide the datasets into classes to find a Maximum Marginal Hyperplane (MMH), where we can easily put the new data point in the correct category in the future. The closest point to hyperplane is referred as Support Vector. Each dataset has a support vector point. The gap between dataset is known as margin. Greater margin will affect better computation result. [8]

C. Decision Tree Classifier

Decision trees are those types of trees which groups attributes by sorting them based on their values. Decision tree is used mainly for classification purpose. Each tree consists of nodes and branches. Each node represents attributes in a group that is to be classified and each branch represents a value that the node can take [9]. An example of decision tree is given in Fig. 3.

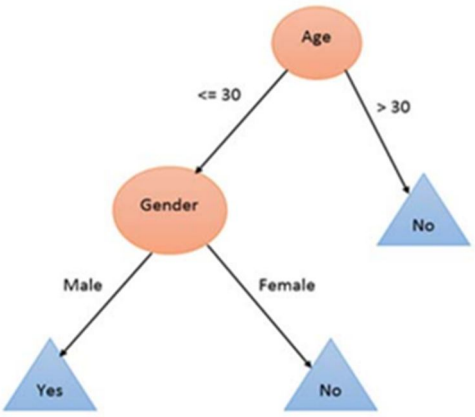


Fig.4 Example of Decision Tree

D. Random Forest Classification

Random Forest is the most common ensemble method, as it consists of a collection of decision trees. The idea behind Random Forest is that we repeatedly select data from the data set and build a decision tree with each new sample and then the most predicted label becomes the class for that data point. The intention of using this classifier is to get a more accurate diagnosis. The reason that

random forest works so well is that a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. The low correlation between models is the key as uncorrelated models predict more accurate results than any of the individual predictions. While some trees may be wrong many others will be right so as a group the trees are able to move in the correct direction. [8]

E. Logistic Regression Model

It is a classification algorithm which is used to predict the chances of happening of a problem. It gives statistical data. It is basically multi-class binary classification. Its concept is mostly similar to the concept of probability i.e., modeling of an event is occurring versus event is not occurring. For e.g., for a hospital dataset, whether the patient is diabetic or not needs to be predicted. Logistic Regression makes use of sigmoid function which takes solution of linear regression and output value between 0 and 1. It has S-shaped curve known as logistic curve.

$$\text{Sigmoid Function} = 1/(1 + e^{-\text{value}})$$

F. XGBoost Classifier

It is short for eXtreme Gradient Boosting package. It is an efficient and scalable implementation of gradient boosting framework by (Friedman, 2001) (Friedman et al., 2000). The package includes efficient linear model solver and tree learning algorithm. It supports various objective functions, including regression, classification and ranking. The package is made to be extendible, so that users are also allowed to define their own objectives easily. [10]

IV. IMPLEMENTATION

The implementation phase has various steps of machine learning and the flow of implementation is shown in Fig.4.

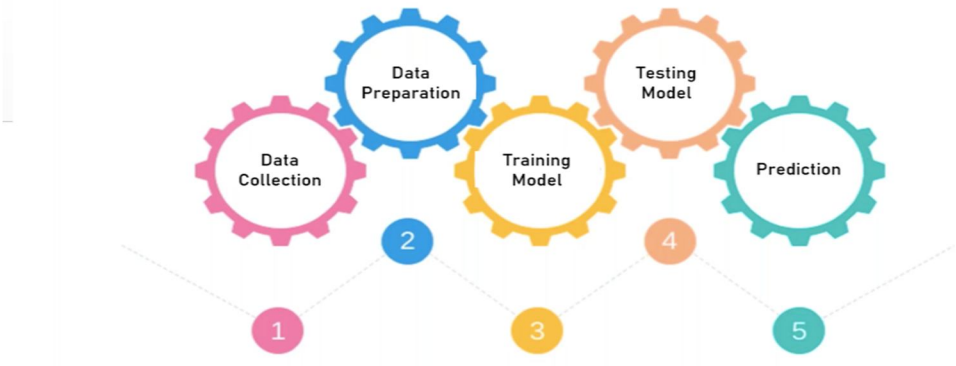


Fig.5 Phases of Implementation

A. Data Collection

All the datasets have been collected from Kaggle community.

1) *Diabetes:* The dataset includes diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

	Number of times pregnant	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Diastolic blood pressure (mm Hg)	Triceps skin fold thickness (mm)	2-Hour serum insulin (mu U/ml)	Body mass index (weight in kg/(height in m)^2)	Diabetes pedigree function	Age (years)	Class variable (0 or 1)
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig.6 Diabetes patients Dataset

2) *Liver Disease:* The data set has been elicit from UCI Machine Learning Repository. This data set contains 416 liver patient records and 167 non liver patient records. The data set was collected from test samples in North East of Andhra Pradesh, India.

	age	gender	tot_bilirubin	direct_bilirubin	tot_proteins	albumin	ag_ratio	sgpt	sgot	alkphos	is_patient
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.90	1
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.89	1
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.00	1
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.40	1

Fig.7 Liver patients Dataset

3) *Chronic Kidney Disease:* This dataset is from UCI Machine Learning Repository. The objective of the dataset is to diagnostically predict whether a patient is having chronic kidney disease or not, based on certain diagnostic measurements included in the dataset.

	Bp	Sg	Al	Su	Rbc	Bu	Sc	Sod	Pot	Hemo	Wbcc	Rbcc	Htn	Class
0	80.0	1.020	1.0	0.0	1.0	36.0	1.2	137.53	4.63	15.4	7800.0	5.20	1.0	1
1	50.0	1.020	4.0	0.0	1.0	18.0	0.8	137.53	4.63	11.3	6000.0	4.71	0.0	1
2	80.0	1.010	2.0	3.0	1.0	53.0	1.8	137.53	4.63	9.6	7500.0	4.71	0.0	1
3	70.0	1.005	4.0	0.0	1.0	56.0	3.8	111.00	2.50	11.2	6700.0	3.90	1.0	1
4	80.0	1.010	2.0	0.0	1.0	26.0	1.4	137.53	4.63	11.6	7300.0	4.60	0.0	1

Fig.8 Kidney patients Dataset

B. Data Preparation

Data preparation includes cleaning the data and replacing missing data with most appropriate data. This step is also called Data Pre-processing. As this paper deals with medical data, any changes have not been made in order to preserve the patient's data.

Data correlation is the way in which one set of data may correspond or relate to another set. The set of correlation values between pairs of its attributes form a matrix which is called a correlation matrix as shown in Fig. 9, Fig. 10 and Fig. 11.

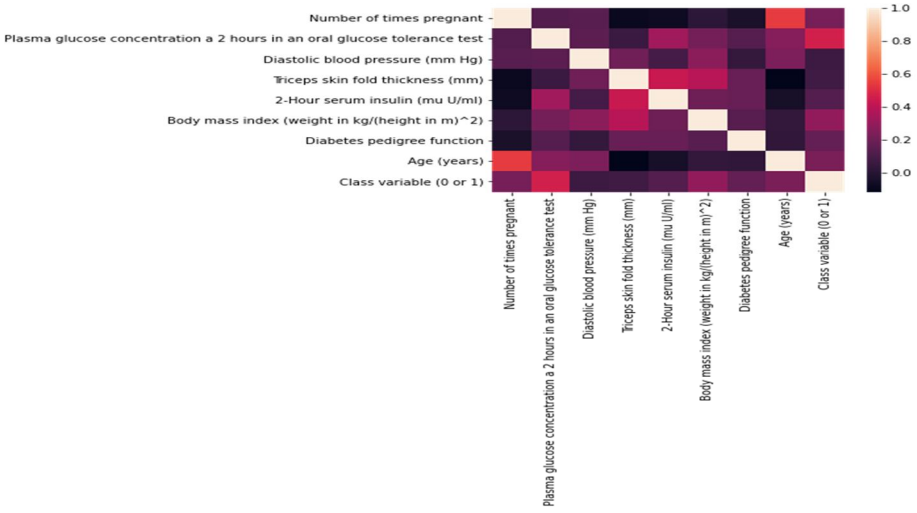


Fig.9 Correlation Heatmap of Diabetes

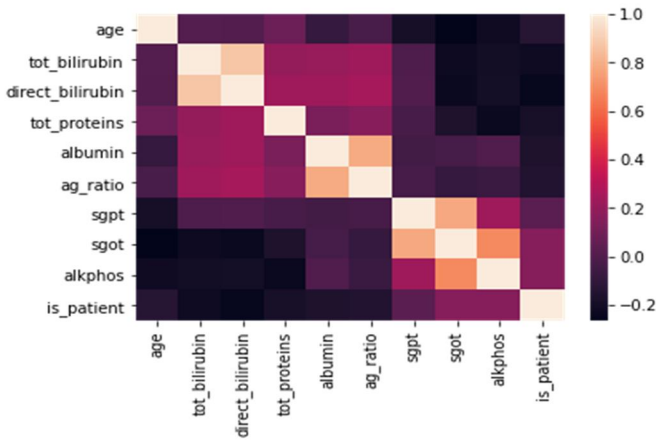


Fig.10 Correlation Heatmap of Liver

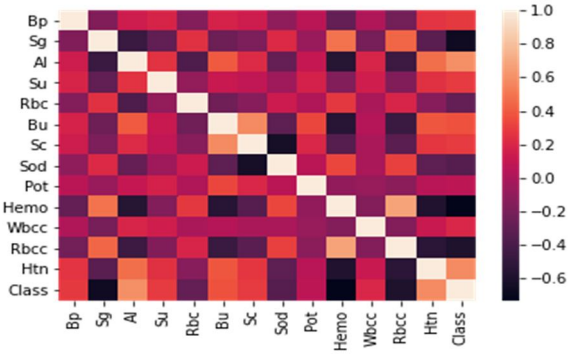


Fig.11 Correlation Heatmap of Kidney

C. Training Model

The datasets are trained using different machine learning algorithms. Every algorithm has different working procedures resulting in varying accuracy. Training is basically the process of giving the machine capability to make further predictions after learning from the training dataset.

```
# XG Boost
from xgboost import XGBClassifier
xgboost_model = XGBClassifier()
xgboost_model.fit(x_train, y_train)
y_pred_xg = xgboost_model.predict(x_test)

xgboost_model.score(x_test , y_pred_xg)*100
```

```
#Logistic Regression Model

from sklearn.linear_model import LogisticRegression
log_model = LogisticRegression()
log_model.fit(x_train, y_train)
y_pred_log = log_model.predict(x_test)
```

```
#Naive Bayes Classification

from sklearn.naive_bayes import GaussianNB
classifier_naive = GaussianNB()
classifier_naive.fit(x_train, y_train)
y_naive_pred = classifier_naive.predict(x_test)
```

```
#Decision Tree Classifier

from sklearn.tree import DecisionTreeClassifier
classifier_dtc = DecisionTreeClassifier()
classifier_dtc.fit(x_train,y_train)
y_dtc_pred = classifier_dtc.predict(x_test)
```

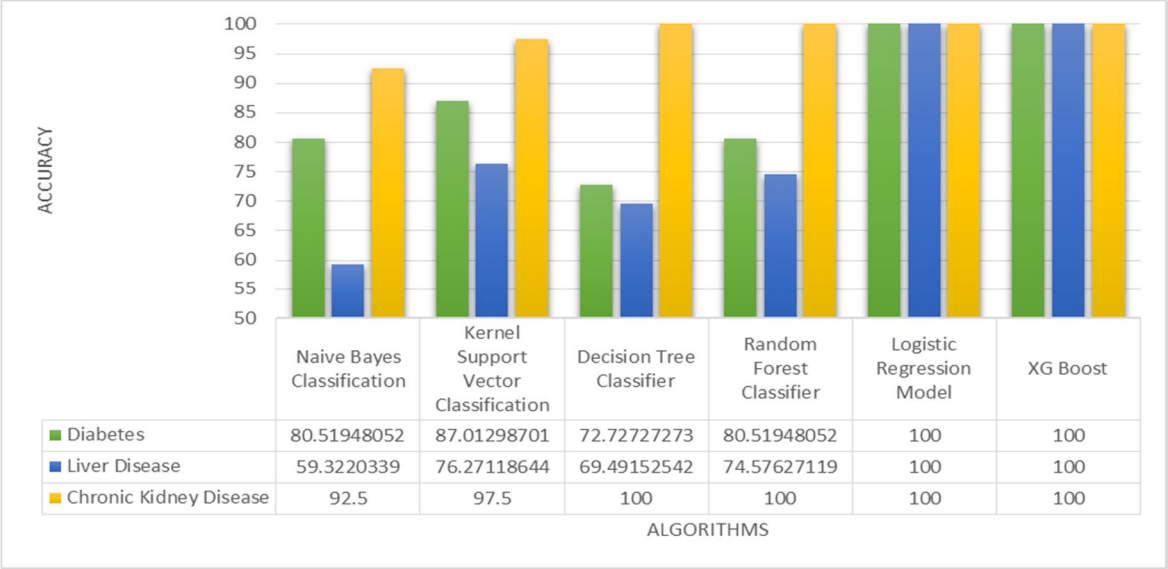
```
#Random Forest Classifier

from sklearn.ensemble import RandomForestClassifier

classifier_rfc = RandomForestClassifier(n_estimators=20, random_state=0)
classifier_rfc.fit(x_train, y_train)
y_rfc_pred = classifier_rfc.predict(x_test)
```

```
#Kernel Support Vector Classification

from sklearn.svm import SVC

classifier_svc = SVC(kernel='linear',random_state=0)
classifier_svc.fit(x_train,y_train)
y_svc_pred = classifier_svc.predict(x_test)
```

Fig.12 Training datasets using different algorithms

D. Testing Model

Testing of a model is done to check the performance of the algorithms in term of accuracy, precision etc. In testing whether the prediction is correct or not is checked using already predefined dataset. Higher the accuracy, better the results.



	Naive Bayes Classification	Kernel Support Vector Classification	Decision Tree Classifier	Random Forest Classifier	Logistic Regression Model	XG Boost
Diabetes	80.51948052	87.01298701	72.72727273	80.51948052	100	100
Liver Disease	59.3220339	76.27118644	69.49152542	74.57627119	100	100
Chronic Kidney Disease	92.5	97.5	100	100	100	100

Fig.13 Algorithms with respect to their accuracy scores in respective disease

E. 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. [11]

In Fig. 14, Fig. 15 and Fig. 16, actual values and predicted values have been compared.

	Class	Predicted Class
0	1	1
1	1	1
2	1	1
3	1	1
4	1	1
...	...	...

Fig.14 Actual vs. Predicted values of Diabetes patients dataset

	is_patient	Predicted is_patient
0	1	1
1	1	1
2	1	1
3	1	1
4	1	1
...	...	...

Fig.15 Actual vs. Predicted values of Liver patients dataset

	Class variable (0 or 1)	Predicted Class variable (0 or 1)
0	1	1
1	0	1
2	1	1
3	0	1
4	1	1
...	...	...

Fig.16 Actual vs. Predicted values of Kidney patients dataset

V. CONCLUSION

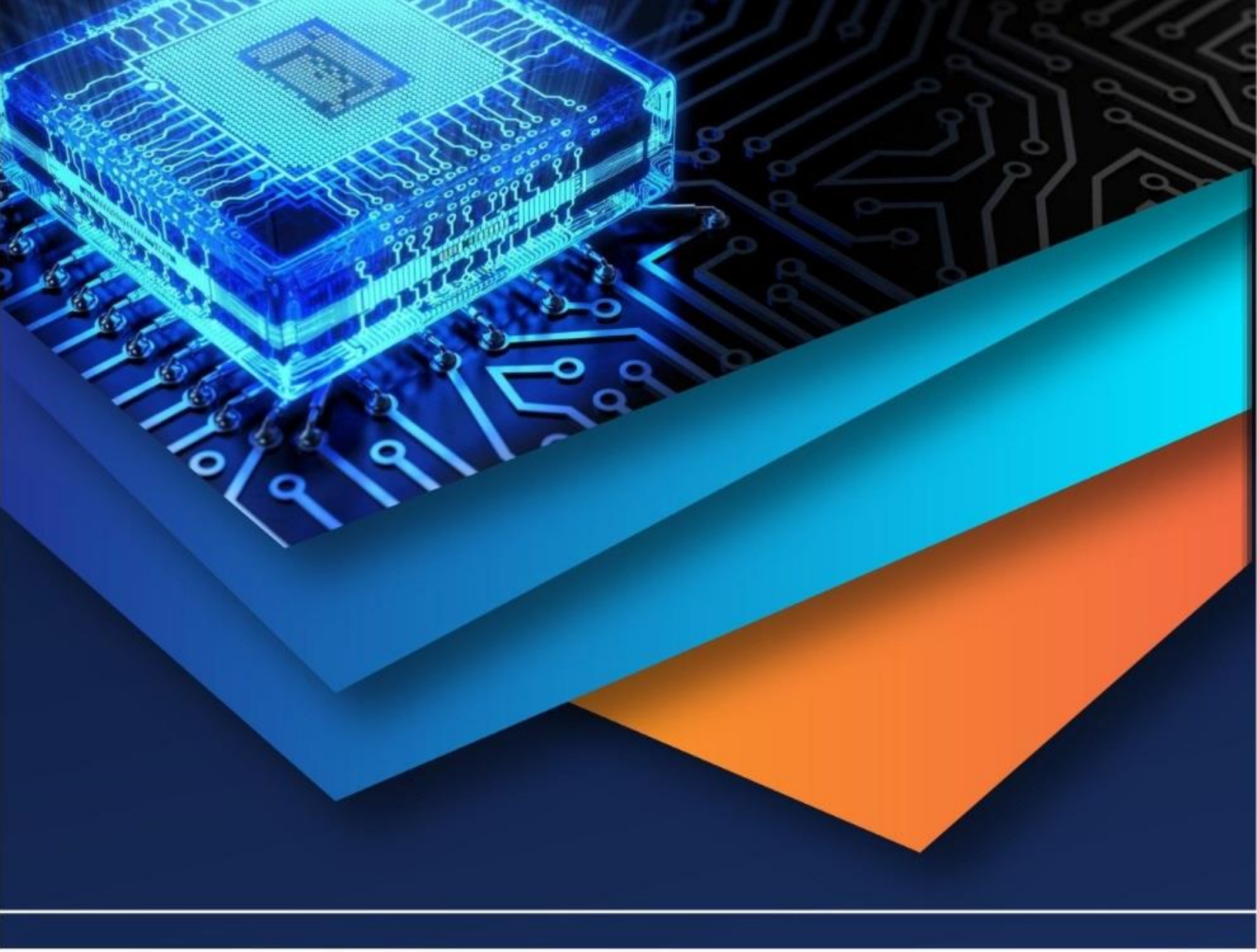
Machine learning has emerged as a field critical for providing tools and methodologies for analyzing the data generated by the biomedical sciences. [8] This review has provided a condensed snapshot of applications of machine learning for the detection of three diseases: Diabetes, Liver Disease and Chronic Kidney Disease using different classifiers of supervised machine learning. Fusion of disparate multimodal and multi-scale biomedical data continues to be a challenge. Further improvements in data can be made like having more features and least null values. Moreover, substantial improvements in Python-based workflow can also be done.

VI. FUTURE SCOPE

Machine learning includes a number of algorithms and techniques to analyze and implement to gain the benefits of them in different fields including healthcare. ML methods can help the integration of computer-based systems in the healthcare environment providing opportunities to facilitate and enhance the work of medical experts and ultimately to improve the efficiency and quality of medical care. ML technologies can be used to identify potential clinical trial candidates, access their medical history records, monitor the candidates throughout the trial process, select best testing samples, reduce data-based errors, and much more. [12]
In future with respect to this model, one can try to develop a system in which most probable disease for a patient can be predicted on the basis of symptoms and moreover test can also be recommended for the predicted disease. A potential future development of the presented work is to apply ML models to other data with different features, concerning the survival prognosis of the patients and early detection of the disease and it can also be developed in web-based application with additional services.

REFERENCES

[1] Diabetes Research Clinical Practice] [Article] https://www.diabetesresearchclinicalpractice.com/article/S0168-8227(19)31230-6/fulltext.
[2] [World Gastroenterology] [Global Burden of Liver Disease-a true burden on Health Sciences and Economies] https://www.worldgastroenterology.org/publications/e-wgn/e-wgn-expert-point-of-view-articles-collection/global-burden-of-liver-disease-a-true-burden-on-health-sciences-and-economies.
[3] [World Kidney Day: Chronic Kidney Disease. 2015] http://www.worldkidneyday.org/faqs/chronic-kidney-disease/.
[4] T. Mitchell, Machine learning, 0-07-042807-7, McGraw Hill (1997), p. 2.
[5] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda,Machine Learning and Data Mining Methods in Diabetes Research, Computational and Structural Biotechnology Journal, Volume 15,2017,Pages 104-116,ISSN 2001-0370.
[6] D. Lowd, P. Domingos, "Naïve Bayes Models for Probability Estimation".
[7] S. Kanchana, "Statistical Analysis Using Machine Learning Approach for Multiple Imputation of Missing Data", in IJRASET, vol.6, February 2018, p.2091
[8] A Comparative Study of Machine Learning Classifiers for Medical Diagnosis, Bhavnath Thakur, Harshit Rohela, Kanishk Gupta, Chhaya Sharma. "A Comparative Study of Machine Learning Classifiers for Medical Diagnosis", Volume 8, Issue IV, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 1748-1752, ISSN: 2321-9653.
[9] S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31 (2007) 249-268.
[10] T. Chen and T. He, "XGBoost: extreme gradient boosting", R Package. Version 0.4-2, 2015.
[11] [DataRobot] [Wiki][Prediction] https://www.datarobot.com/wiki/prediction/.
[12] [Upgrad] [Blog][Machine Learning Applications in Healthcare] https://www.upgrad.com/blog/machine-learning-applications-in-healthcare/.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY