International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 01 | Jan 2022 www.irjet.net p-ISSN: 2395-0072

HealthOrzo – Your Health Matters

Vaishali Jabade[1], Sarthak Bhake[2], Sahil Parekh[3], Sakshi Kulkarni[4], Sejal Sayam[5], Shajjad Shaikh[6]

[1]Assistant Professor, Vishwakarma Institute of Technology, Pune, Maharashtra, India
[2-6]Students, Dept. of Electronics and Telecommunications, Vishwakarma Institute of Technology, Pune

---

Abstract - *Health is Wealth. This Sentence is becoming more and more accurate nowadays. People are becoming more and more health conscious and want to be in the best shape possible. Our Project caters to the same need of the People. Artificial Intelligence and Machine Learning have done wonders in the healthcare industry by using the classification and regression algorithms to predict diseases based on the data given by the user. The Project Predicts 4 serious diseases that is Diabetes, Heart, Kidney and Liver Ailment using different supervised Machine Learning Algorithms Like Support Vector Machines, Logistic Regression etc. The Algorithms are trained and tested on datasets which are taken from Kaggle for the different diseases that the Project Predicts. All the 4 machine learning models are integrated in a user-friendly website using the flask library at the backend. HealthOrzo is thus a website having all 4 prediction algorithms at 1 single place in the form of a website.*

Key Words: **Machine Learning, Supervised Machine Learning, Support Vector Machine (SVM), Logistic Regression, Random Forest, Python, Flask, Health Prediction.**

1.INTRODUCTION

During these Difficult Times Health is utmost Importance to Everyone. People are more conscious about their Health now more than ever. The pandemic has exposed of how much havoc can be caused by things which we have been neglected for years. Specially the attention was drawn to comorbid diseases which played a important role in the pandemic. The comorbid diseases can be termed as diseases which alter your entire lifestyle and can impact your future lifestyle too. In short these are the ones which need to be taken seriously. Some of The Examples being Diabetes, Heart Disease, Chronic Kidney Disease, Liver Disease, Lung Failure etc. Moreover, the people which are having these diseases were given special priority when the covid vaccination drives were conducted. This just shows how important these diseases are how they should be taken very seriously. Now, the question arises that how can we detect or predict if a particular person has this disease at present or may have it in the Future. This is where the domain of Machine Learning and Data Science comes in. Now that we have sufficient introduction about comorbid diseases lets shift our focus towards the introduction of Machine Learning and its related terms which are extensively used nowadays in the industry.

2. INTRODUCTION TO MACHINE LEARNING

Machine Learning or Popularly abbreviated as ML is a technique or process in which the machine learns by itself with the help of different computer algorithms which are trained and tested on data. It turns applies the same approach and algorithm on the data again without needing any further human intervention. This is as simple the definition of Machine Learning can get. ML as a domain has 3 different Subparts to it namely, Supervised Learning, Unsupervised Learning and Re-enforcement Learning. All these Subdomains have some Algorithms which fall under them, to understand it better please refer the figure below.

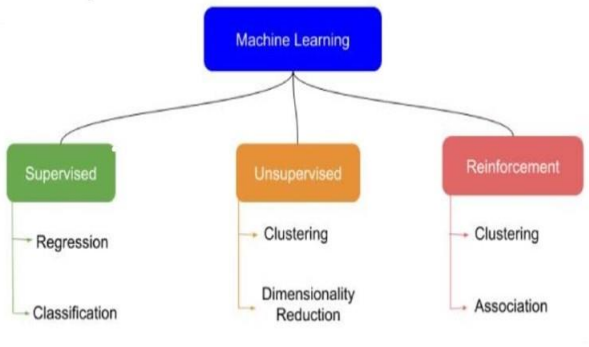

Figure 1- Sub Domains of Machine Learning

As you can see the distinction between the 3 sub domains, let us understand which all different Algorithms fall under these sub domains.

Supervised Machine Learning

It is defined as the domain where everything happens in a guided path, being analogous to be taught by a teacher. In this the algorithms are trained well on train data and are then tested on the training data. This sub domain is used for the classifications and prediction based on the data. Some of the different algorithms which fall under supervised machine learning are:

- Linear Regression.
- Logistic Regression.
- Support Vector Machines (SVM).
- Random Forests.

- Decision Trees.
- Naïve Bayes. etc.

Unsupervised Machine Learning

As you must have probably guessed by its name, the algorithms which fall under this domain are performed on data that is not supervised or labeled. The models which are built using unsupervised algorithms are self-sufficient to find patterns and insights from the data on their own. This being analogous to how our human brain learns new things and finds patterns/draws insights form a given particular thing. Some of the different algorithms which fall under unsupervised machine learning are:

- K-means Clustering.
- Hierarchical Clustering.
- K-nearest Neighbors (Knn). Etc.

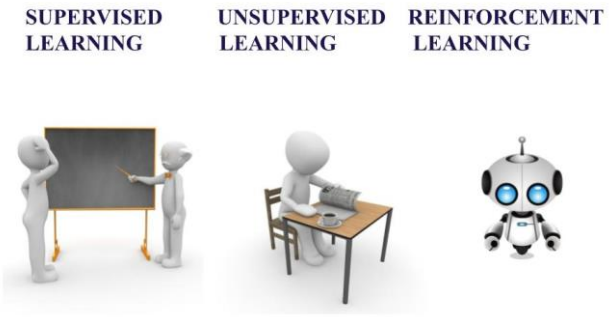

Figure 2 – Difference between the 3 subdomains in a visual format

3. RELATED WORK

During the research of this Project, we scrapped many research papers in order to study different approaches and algorithms which would suit our problem statement in the best way. As mentioned earlier we are going to predict 4 different comorbid diseases like diabetes, Heart, kidney and liver disease. In order to build 4 different Machine Learning Models each catering to a particular disease we had to go through different papers in order to choose the optimal algorithms with the best accuracy and the datasets which suit those particular algorithms. While doing Literature survey for heart disease prediction we had to lock upon which are the key parameters which we need to consider for the prediction. A. Gavhane et.al [1] recently reported and mentioned the different features necessary for heart disease prediction, which being Body Mass Index, Heart Rate, Fasting Blood Sugar, Age, Cholesterol etc. After locking in on the parameters required, we finalized the dataset we have used for heart prediction courtesy of A. Singh et.al [2] where they had specified the dataset and elaborated on each feature. Now the dataset being finalized for heart we had to choose the algorithm A. H. Chen et.al [3] had implemented Random Forest Classifier and had an 80% accuracy but opted for logistic regression as our algorithm and got 83% accuracy. Now coming to the Kidney Disease Prediction model, G. Chen et al [4] elaborated on the parameters which are considered during kidney disease and prediction and implemented the prediction using neural networks. Now as the parameter are fixed, we fixed our dataset which was elaborated by P. Chittora et.al [5]. Diabetes was also a very key Model in our Project and the dataset was already decided but we now had to choose the algorithm. A. Anand et.al [6] implemented random forest classifier but the accuracy on our data set was quite low for this particular algorithm. A. Mir et.al [7] in their research showed that most of the people have opted Support Vector Machine (SVM) as their optimal algorithm for diabetes prediction, hence we opted for the same algorithm. S. Sontakke et.al [8] in their paper illustrated about the classification and prediction of the model. Thus, we decided random forest as our algorithm for our liver model.

This was all for the Literature Survey now lets' have a look and what is the methodology and different steps which are involved in the building of these 4 Machine Learning Models.

4. METHODOLOGY

While Building Machine Learning Models there some steps/methods which are carried out every time. You can also call them as SOP Standard Operating Procedure while building Machine Learning Models. Let's have a Look at all of them one by one.

A. Importing the Libraries

First of all, we need to import the necessary libraries and tools needed for our project. Be it some pre-processing tools or ML Enabling Libraries. These are the Following Libraries/Frameworks we have used/imported for our Project.

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sci-kit Learn

B. Data Pre-processing and Visualization

This is a very key step towards to building of our model. This is the step where all our data gets cleaned and visualized. After the data is visualized using different charts and metrics it is easy to draw insights from charts that are presented. There are Different Types of Plots which are useful, some of them are.

- Histogram.
- Scatter Plot.
- Box Plot.
- Heat Maps.
- Pie Charts.
- Distribution Plot.

Figure 3- Different Types of Charts and Visualization Techniques

C. Train Test Split

This is step where we split our dependent variables and independent variable into two different series or data frames and then we perform the train test split where the divide the data set into four variables namely, X_test, y_train, X_test, y_test . We import the train test split call from a library named sk-learn which has in built instances for this particular split. You can refer the syntax given in the figure attached below.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=10)
```

Figure 4 – Train Test Split using Sci-kit Learn

D. Saving the Model

After we have successfully completed building our model it's time to save it in the form of a file for further use. To save these particular models we have used pickle library which enables swift saving of models and the loading it anywhere as file with a. pkl extension. We have taken the full advantage of this property while developing our flask app where we render a particular model based on the input given by the user.

```
import pickle

pickle.dump(classifier, open("diabetes.pkl",'wb'))
```

Figure 5 – Saving our model in a pickle model.

5. IMPLEMENTATION

Our Project has been divided into 2 parts, one is the model folder and other is the web part where we have the HTML Templates and the Flask app running in the Back end of the code. We have used 4 Models namely to cater to our diseases that is Heart, Diabetes, Liver and Kidney. We'll Focus one by one on all our models.

A. Diabetes Model

For the Diabetes Model we have imported the dataset which has the following features attached in the Figure below.

Pregnancie	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesPe	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1

diabetes

Figure 6 – Dataset used for Diabetes Prediction

A.1 Algorithm

To predict the diabetes, we have used Support Vector Machine (SVM) Algorithm. It is an Algorithm which comes under supervised machine learning and this algorithm is mainly used for classification purposes. The main Goal of SVM is to create a best line or somewhat like a boundary that can segregate n dimensional Place into classes so that we can segregate future data into the correct category. The boundary which segregates the data points is called Hyperplane.

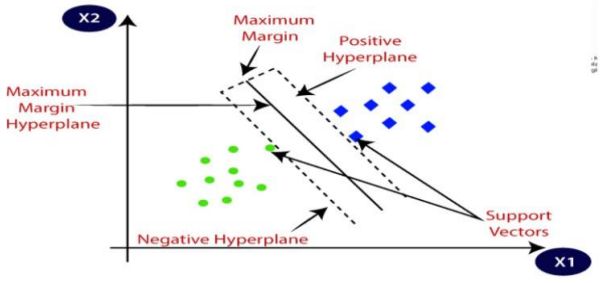

Figure 7 – Support Vector Machine Representation

A.2 Accuracy Score

The score which represents the performance of the algorithm is called the Accuracy score. We have Achieved 75% Accuracy on this model by using SVM.

```
X_test_prediction = classifier.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, y_test)
```

```
print('Accuracy score of the test data : ', test_data_accuracy*100)
```

```
Accuracy score of the test data :  75.19685039370079
```

Figure 8 – Accuracy Score of Diabetes Model

B. Heart Model

The dataset which is used for this model has various different features which are influential in the prediction. The features can be recalled from the figure attached below.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
48	0	2	130	275	0	1	139	0	0.2	2	0	2	1

Heart

Figure 9 – Dataset used for Heart Disease Prediction

B.1 Algorithm

In order to predict this particular disease, we have used Logistic Regression as our Algorithm. This Particular Algorithm predicts the output of the dependent variable. Thus, the Output should also be a discrete value. It returns a probabilistic value which has a value between 0 to 1. The mathematical representation of logistic regression can be found in the figure attached below.

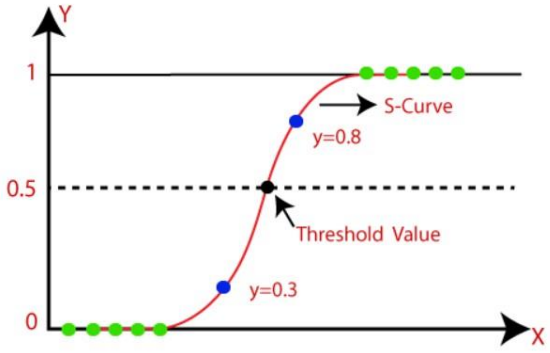

Figure 10 – Logistic Regression Representation

B.2 Accuracy Score

The Accuracy Score of this model after using Logistic Regression 81%. The accuracy testing is done on both train data and test data.

```
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, y_test)
```

```
print('Accuracy on Test data : ', test_data_accuracy*100)
```

```
Accuracy on Test data :  81.9672131147541
```

Figure 11– Accuracy Score of Heart Model

C. Liver Model

The liver Model is where we first classify and then predict the disease based on the dataset which is being used. The dataset has different features and pigment ratios which are key in assessing whether a person has a fit liver or not. These Ratios not only signify one's liver's well-being but also signify the overall health of a person. The features can be seen from the image attached below.

Age	Gender	Total_Bilir	Direct_Bili	Alkaline_P	Alamine_A	Aspartate_	Total_Prot	Albumin	Albumin_a	Dataset
65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
58	Male	1	0.4	182	14	20	6.8	3.4	1	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
26	Female	0.9	0.2	154	16	12	7	3.5	1	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	2
55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
72	Male	2.7	1.3	260	31	56	7.4	3	0.6	1

liver

Figure 12 – Dataset Used for Liver Disease Prediction

C.1 Algorithm

For this Prediction we have used Random Forest classifier and regressor. This particular algorithm comes under ensemble learning. The Algorithm uses multiple decision tress to classify and predict the outcome of a particular situation. The nominal Working of Random Forests can be seen in the Figure below

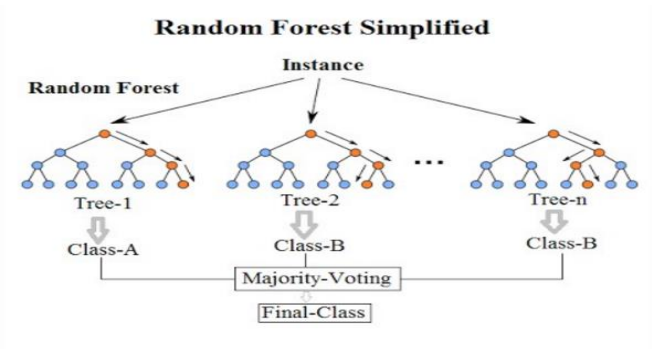

Figure 13 – Random Forest Classifier

C.2 Accuracy Score

The liver model after applying the random forest algorithm gives an accuracy of 73% on the testing dataset.

```
print(f"Accuracy is {round(accuracy_score(y_test, model.predict(X_test))*100,2)}")
```

```
Accuracy is 73.58
```

Figure 14 – Accuracy Score of Liver Model

D. Kidney Model

The kidney model and dataset are one most complicated yet the most interesting case. The dataset has 26 different columns containing various features used in the Machine Learning Model. The dataset can be seen in the figure attached below.

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo
0	0	48	80	1.02	1	0		normal	notpresen	notpresen	121	36	1.2			15.4
1	1	7	50	1.02	4	0		normal	notpresen	notpresent		18	0.8			11.3
2	2	62	80	1.01	2	3	normal	normal	notpresen	notpresen	423	53	1.8			9.6
3	3	48	70	1.005	4	0	normal	abnormal	present	notpresen	117	56	3.8	111	2.5	11.2
4	4	51	80	1.01	2	0	normal	normal	notpresen	notpresen	106	26	1.4			11.6
5	5	60	90	1.015	3	0			notpresen	notpresen	74	25	1.1	142	3.2	12.2
6	6	68	70	1.01	0	0		normal	notpresen	notpresen	100	54	24	104	4	12.4
7	7	24		1.015	2	4	normal	abnormal	notpresen	notpresen	410	31	1.1			12.4
8	8	52	100	1.015	3	0	normal	abnormal	present	notpresen	138	60	1.9			10.8
9	9	53	90	1.02	2	0	abnormal	abnormal	present	notpresen	70	107	7.2	114	3.7	9.5
10	10	50	60	1.01	2	4		abnormal	present	notpresen	490	55	4			9.4
11	11	63	70	1.01	3	0	abnormal	abnormal	present	notpresen	380	60	2.7	131	4.2	10.8

Kidney

Figure 15 – Dataset used for Kidney Disease Prediction

D.1 Algorithm

The Algorithm used for this is same as the Liver model that is Random Forest Classifier and Regressor. The theory is same as mentioned earlier in the paper.

D.2 Accuracy Score

The Accuracy score for the Kidney model is inch perfect. The model after being implemented in the Random Forest Classifier gives a 100% Accuracy.

```
confusion_matrix(y_test, model.predict(X_test))
```

```
array([[23,  0],
       [ 0,  9]], dtype=int64)
```

```
print(f"Accuracy of the model is={(accuracy_score(y_test, model.predict(X_test))*100)}%")
```

```
Accuracy of the model is=100.0%
```

Figure 16 – Confusion Matrix and Accuracy Score of Kidney Model

6. RESULTS

The results of the ML models are rendered on a front-end user interface. In our Project we have created a simple static website which will serve as a presentable interface for our project. The website contains a landing page and 4 other pages catering to diseases Diabetes, Heart, Kidney and Liver respectively. The 4 pages contain html forms which take the input from the user and pass those inputs to the models with the help of Flask app which is named as app.py. The prediction is basically in 2 categories that is whether you have the disease or you do not have the disease. This is computed by a function we have defined in the flask app which passes the user input as a numpy array to the ML Model and returns the output as the prediction.

```
def predict(values):
    if len(values) == 8:
        to_predict = np.array(values).reshape(1, 8)
        loaded_model = pickle.load(open("models/diabetes.pkl", "rb"))
        result = loaded_model.predict(to_predict)
        return result[0]

    elif len(values) == 13:
        to_predict = np.array(values).reshape(1, 13)
        loaded_model = pickle.load(open("models/heart.pkl", "rb"))
        result = loaded_model.predict(to_predict)
        return result[0]

    elif len(values) == 18:
        to_predict = np.array(values).reshape(1, 18)
        loaded_model = pickle.load(open("models/kidney.pkl", "rb"))
        result = loaded_model.predict(to_predict)
        return result[0]

    elif len(values) == 10:
        to_predict = np.array(values).reshape(1, 10)
        loaded_model = pickle.load(open("models/liver.pkl", "rb"))
        result = loaded_model.predict(to_predict)
        return result[0]
```

Figure 17 – App.py file

In the figure above you can see that we are loading our saved pickle models and rendering them based on the length of the user input. This is a very logical and effective way of managing the models and making sure the correct model is loaded according to the user input.

```
@ app.route("/predict", methods=['GET', 'POST'])
def predictPage():

    if request.method == 'POST':
        to_predict_list = request.form.to_dict()
        to_predict_list = list(to_predict_list.values())
        to_predict_list = list(map(float, to_predict_list))
        pred = predict(to_predict_list)

        return render_template('predict.html', pred=pred)
```

Figure 18 – App.py file

In the Figure Above it is the function which accepts the HTML Form input from the respective templates and typecasts it into a list and changes its datatype to float. After this is done the variable in which the input is stored is then passed to the function which has the Machine Learning Model loaded. (The function showed in Figure 17)

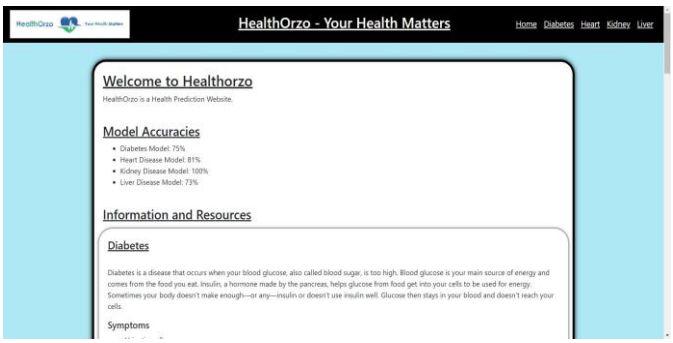

Figure 19 – Landing Page of the Website

In the above Figure you can see the landing page of our website. Here we have displayed various resources and information regarding diseases that we predict. We have included several blog links and embedded videos to educate the users about Diabetes, Kidney, Liver and Heart Diseases.

Figure 20 – Diabetes Prediction Form

In the Figure attached above, there is a form which accepts the user input and passes it to the flask file viz. app.py. After the user has given input to the website the input is processed and the result/outcome is displayed.

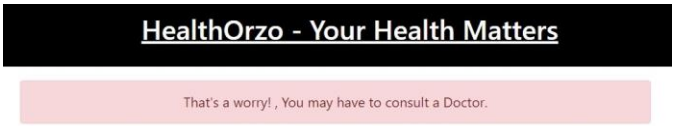

Figure 21 – Output Showing the user has the Disease.

Figure 22 – Output Showing that the user is Healthy

The above 2 Figures that is Figure 21 and 22 are the output windows. This is what the user sees as an output after the input is processed. The output is in the form of a message just to keep it as simple as possible.

7. CONCLUSION

From this we can conclude that all the 4 machine learning models are working properly while being integrated in a website. The users can give their input and successfully obtain the result/prediction after pressing the submit button. The main takeaway point from this Project is the way in which technologies like Machine Learning can predict and have influence in our day to day lives. A few years Ago, it would've been difficult to predict health of people considering there is so much at stake. But thanks to amazing research and technology that we have at our disposal today it is done very easily. It is always a pleasure to learn new skills and implement them on a project that adds some value to betterment and the well-being of the society.

ACKNOWLEDGEMENT

We would like to thank our College Vishwakarma Institute of Technology, Pune for providing us with a platform to conduct our research and implement our project. We would like to also express gratitude to our Project guide Prof. Dr. Vaishali Jabade ma'am for her constant guidance and support to us.

REFERENCES

[1] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.

[2] A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[3] A. H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng and E. J. Lin, "HDPS: Heart disease prediction system," 2011 Computing in Cardiology, 2011, pp. 557-560.

[4] G. Chen et al., "Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 100497-100508, 2020, doi: 10.1109/ACCESS.2020.2995310.

[5] P. Chittora et al., "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," in IEEE Access, vol. 9, pp. 17312-17334, 2021, doi: 10.1109/ACCESS.2021.3053763.

[6] A. Anand and D. Shakti, "Prediction of diabetes based on personal lifestyle indicators," 2015 1st International Conference on Next Generation Computing Technologies (NGCT), 2015, pp. 673-676, doi: 10.1109/NGCT.2015.7375206.

[7] A. Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697439.

[8] S. Sontakke, J. Lohokare and R. Dani, "Diagnosis of liver diseases using machine learning," 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), 2017, pp. 129-133, doi: 10.1109/ETIICT.2017.7977023.