Disease Prediction System

Ms. Pooja G
Associate Professor
Department o f ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

Harish J
B. TECH ARITIFICIAL INTELLIGENCE AND DATA SCIENCE – 2nd YEAR

Anupama R
B. TECH ARITIFICIAL INTELLIGENCE AND DATA SCIENCE – 2nd YEAR

Krish Joyeaal A B
B. TECH ARITIFICIAL INTELLIGENCE AND DATA SCIENCE – 2nd YEAR

ABSTRACT: The combination of diabetes and heart disease represents a significant threat to global health, emphasizing the importance of early disease prediction for effective treatment. In this regard, there are challenges in analysing clinical data, but data science techniques offer promising solutions. Advances in healthcare technology, especially machine learning algorithms, allow valuable insights to be gained from large medical data. Integrating these algorithms into predictive models provides a new approach to disease diagnosis. In one such case, a predictive model for heart disease and diabetes was developed using a support vector classifier and logistic regression algorithms. In addition, a user-friendly graphical user interface has been created that makes it easy for medical professionals to enter patient information and obtain accurate forecasts. This acceleration includes the potential to streamline healthcare decision-making processes and improve patient outcomes through timely intervention.

Keywords: Diabetics, Heart disease, Machine learning, Logistic Regression, SVM

INTRODUCTION

Modern fast-paced culture is seeing an increase in the prevalence of diabetes and heart disease due to factors like unhealthy eating habits, high levels of stress, and sedentary activity. Processed meals, ongoing stress, lack of sleep, and sedentary employment are risk factors. Pollution and other environmental issues of modern living are important. People should prioritize stress relief, a balanced diet, regular exercise, and adequate sleep. They should also be aware of the effects that their lifestyle decisions have on them. It is essential that you receive qualified medical advice in order to treat these health issues and identify them early. Heart disease and diabetes are on the rise as a result of lifestyle choices like eating poorly, sitting a lot, and being under a lot of stress. Distinguishing these circumstances early and overseeing them successfully are fundamental for limiting their effect on people's wellbeing results. With headways in innovation and information examination, there is a developing an open door to foster creative answers for anticipating and forestalling sicknesses. Because of this critical need, the improvement of a heart and diabetes infection forecast framework arises as a promising system to recognize early advance notice signs, tailor risk evaluations, and work with ideal mediations. By bridling AI calculations, this framework expects to engage people to make proactive strides in dealing with their wellbeing, while additionally empowering medical services suppliers to convey more customized and proficient consideration. This paper digs into its part in upgrading early location and the executives of coronary illness and diabetes.

LITERATURE REVIEW

In modern healthcare, disease prediction systems for diabetes and heart disease have become indispensable tools for identifying individuals at risk and implementing preventive measures. These systems utilize cutting-edge technologies, such as wearables, electronic health records, and machine learning algorithms, to scrutinize a diverse range of information and furnish personalized risk assessments. The primary objective is to shift healthcare focus from reactive to proactive and preventive approaches.

Machine learning techniques hold a significant role in prediction, and a variety of methodologies have been employed in recent research to develop prediction models for these diseases. Due to their proficiency in analysing intricate datasets, machine learning algorithms such as neural networks, random forests, and support vector machines are frequently employed. The integration of genetic data, wearable technology, and electronic health records has significantly improved the precision and accuracy of these systems, allowing for a more comprehensive evaluation of individual health profiles. Although

traditional factors such as age, family history, and lifestyle choices remain significant, emerging research emphasizes the importance of novel biomarkers such as genetic markers, inflammatory markers, and metabolic indicators in enhancing the sensitivity and specificity of prediction models. Technological advancements have enabled real-time monitoring of health metrics, leading to the development of dynamic prediction systems that allow for timely interventions and personalized recommendations. The integration of remote monitoring into prediction models has shown promise in enhancing the effectiveness of preventive strategies and improving patient outcomes. While various machine learning techniques have been used to predict disease outcomes using diverse datasets, this review focuses specifically on the application of the Support Vector Classifier and Logistic Regression. Recent research demonstrates that patient involvement and empowerment in prediction systems are important. This emphasis is placed on the positive impact of personalized lifestyle interventions, including dietary modifications, exercise regimens, and stress management, on reducing the risk of heart disease and diabetes.

PROPOSED METHODOLOGY

The suggested methodology uses a multimodal approach that incorporates different technologies and approaches to accurately identify persons at risk of acquiring diabetes and heart disease. First, the system analyses a wide range of patient data by using advanced data analytics methods, particularly machine learning algorithms, to identify patterns and trends that suggest the risk of diabetes and heart disease. Among these algorithms are the logistic regression and support vector classifiers. Through the integration of many data sources and the application of sophisticated analytics tools, the suggested methodology seeks to provide a thorough and individualised solution for illness prediction and prevention. Through the development of an intuitive graphical user interface, the user may interact with the system. The methodology of the heart and diabetes prediction system is centred on proactive risk identification, tailored intervention, and ongoing monitoring to enhance early detection and management of these disorders, which will ultimately improve individual health outcomes.

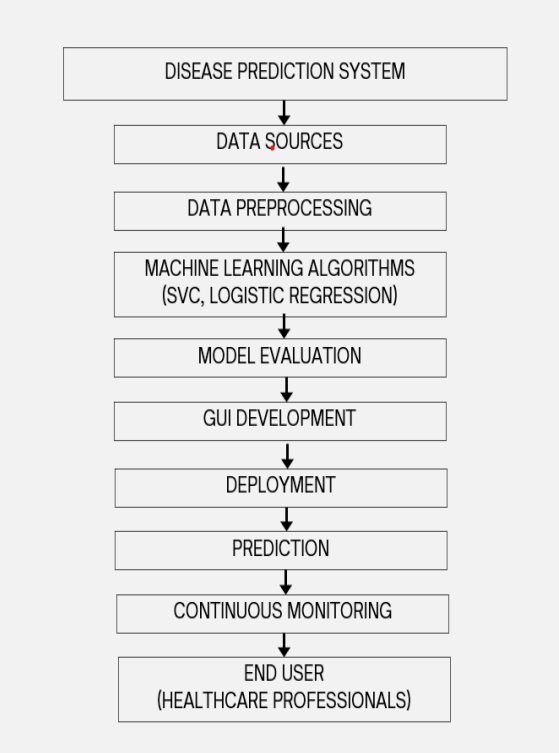

Fig.1 Proposed Methodology

DATA COLLECTION

Gather a variety of datasets from reliable sources, making sure they cover a diversity of lifestyles, regions, and demographics. This will offer a thorough sample for creating a prediction model that is representative of a large population. Use consistent procedures for gathering data to reduce biases and preserve consistency across sources.

DATA PRE-PROCESSING

To preserve important information, handle missing values with suitable techniques like mean imputation or predictive imputation. To keep outliers from adversely influencing the predictive model and maintain the model's robustness, recognize and handle them.

FEATURE SELECTION

To guide the selection process, conduct exploratory data analysis to comprehend the links between prospective characteristics and goal variables.

MODEL-SELECTION

To improve accuracy and generalization, assess how well different machine learning algorithms—including ensemble methods—perform. Examine the assumptions and constraints of every model and choose the ones that most closely match the features and the objectives of the predictions.

MODEL TRAINING

To evaluate performance, train models with subsets of the data, noting efficiency and accuracy metrics. Models should be regularized to reduce overfitting and enhance their capacity to generalize to new data.

PREDICTION

Employ the algorithms that have been taught to compute the likelihood of each category and forecast the existence of illnesses. Depending on the class with the highest likelihood, the output will show whether the disease is present (1) or absent (0).

MODEL EVALUATION

Examine different evaluation measures while keeping in mind the unique requirements of disease prediction, such as accuracy, precision, recall, and F1-score. To find out how modifications to model parameters impact prediction results, perform sensitivity analysis.

GUI DEVELOPMENT

To make sure the GUI design process satisfies user expectations, incorporate user feedback. Use responsive design concepts to make the user interface (GUI) as efficient as possible across a range of devices.

DEPLOYMENT AND MONITORING

Deployment and Monitoring

Create a solid deployment pipeline that includes rollback and version control features. To enable prompt interventions, set up methods for tracking model drift, user engagement, and system performance.

SYSTEM DESIGN

Our disease prediction system is primarily concerned with receiving user input, processing it to extract specific insights, and producing a result. In that instance, we particularly employ logistic regression to predict heart disease and support vector classifier to predict diabetic disease.

SUPPORT VECTOR CLASSIFIER

Support vector machines (SVMs) continue to be effective in high-dimensional environments even when there are more dimensions than samples. Their memory efficiency stems from their utilization of support vectors, which are a subset of training points. Because of its versatility, SVMs can be used to make decisions using a variety of kernel functions. They can be outlier-sensitive and inherently support binary and multi-class classification. Even with this sensitivity, its effects can be lessened with careful kernel selection and parameter optimization. SVMs work best on complex, small datasets. First, logistic regression is frequently attempted; if it is unable to yield accurate results, SVM without a kernel, or linear SVM, can be a good substitute. Depending on the characteristics of the dataset, the performance of both approaches is comparable.

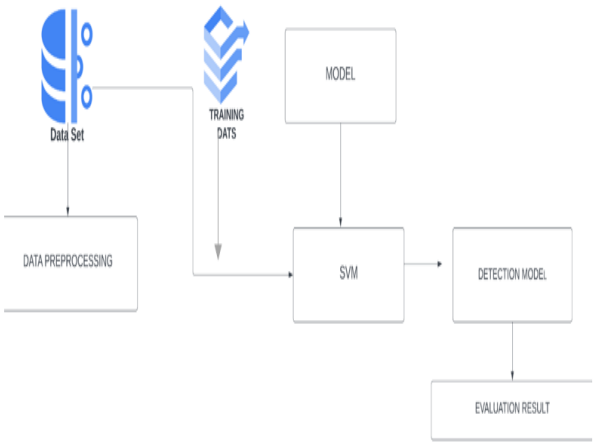

Fig.2 Flowchart for SVM

LOGISTIC REGRESSION

A data analysis method called logistic regression is used to determine the correlations between two elements and forecast one's value based on the other. It is extensively employed in machine learning and artificial intelligence. Organizations can use insights from their data for predictive analysis, cost reduction, efficiency gains, and faster scaling when they use machine learning models constructed with logistic regression. An essential part of logistic regression is the logistic function, often known as the logit function, which maps one variable as a sigmoid function of another. Logistic regression presupposes a linear relationship between variables when more than one influences the dependent variable. The logistic regression formula can be used to calculate the final output variable: \[y = f(\beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_nx_n)\]

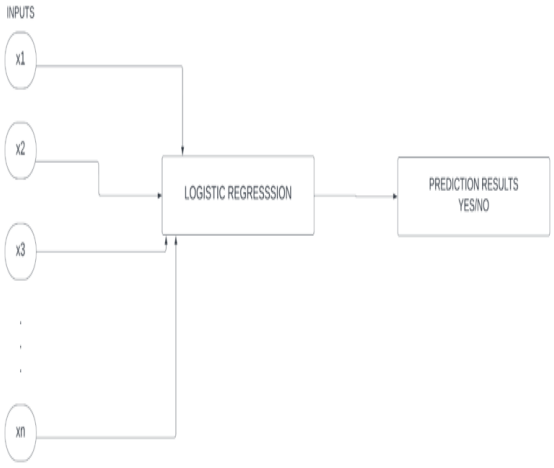

Fig.3 Flowchart for logistic regression

FLOW CHART

A flowchart for a heart disease and diabetes prediction system would outline the steps involved in predicting whether an individual is at risk for either condition. Here's a general description of such a flowchart:

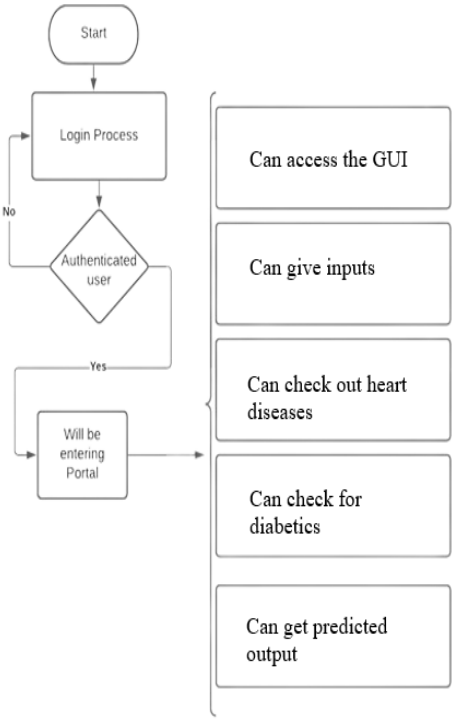

Fig.4 Flowchart of the system

A flowchart for a diabetes prediction system would outline the steps involved in predicting whether an individual is at risk for either condition. Here's a general description of such a flowchart:

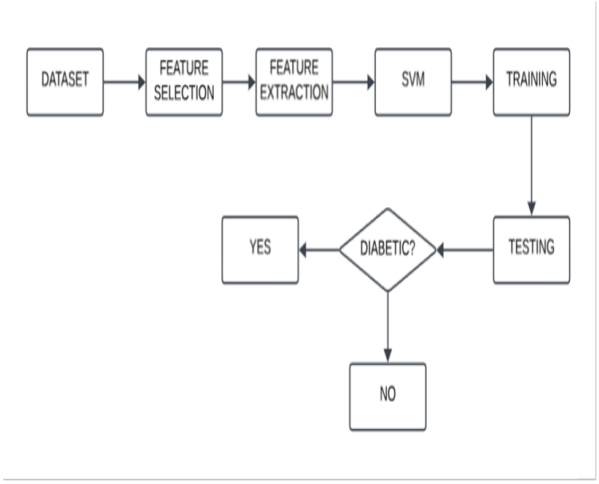

Fig.5 Flowchart of the diabetics system

A flowchart for a heart disease prediction system would outline the steps involved in predicting whether an individual is at risk for either condition. Here's a general description of such a flowchart:

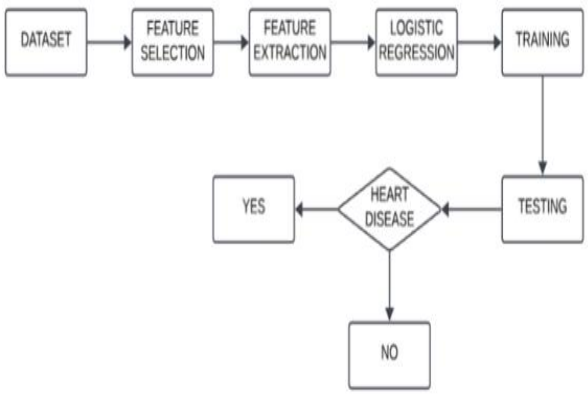

Fig.6 Flowchart of the heart disease system

DATASET

The dataset includes clinical and demographic characteristics, among other heart health-related information. Factors such as age, gender, kind of chest pain, blood pressure at rest, serum cholesterol, maximal heart rate reached, angina triggered by exercise, and more. The target variable (binary classification: 0 for absence, 1 for presence) shows whether or not cardiac disease is present.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
22	59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
23	44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
24	42	1	0	140	226	0	1	178	0	0	2	0	2	1
25	61	1	2	150	243	1	1	137	1	1	1	0	2	1

Fig.7 Dataset for heart

	Pregnancie	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesPe	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90	0	0	39.8	0.451	41	1
25	9	119	80	35	0	29	0.263	29	1

Fig.8 Dataset for diabetics

IMPLEMENTATION

Step 1: Download appropriate Kaggle datasets with pertinent health data, such as demographics, medical history, and outcomes of diabetes and heart disease diagnostic tests.

Step 2: To ensure consistency, normalize numerical features and use methods such as mean or median imputation to handle missing data. Use techniques such as one-hot encoding when working with categorical data.

Step 3: Consider feature interactions that impact health outcomes and add other features, such as BMI and cholesterol ratios, to capture nuanced health indicators.

Step 4: Try other machine learning methods for diabetes and heart disease prediction, such as SVM, Logistic Regression, and Decision Trees. To ensure generalizability and optimize model parameters, apply cross-validation.

Step 5: Evaluate model performance with healthcare-specific metrics including sensitivity and specificity, as well as metrics like ROC-AUC, precision-recall curves, and confusion matrices.

Step 6: Design a user-friendly layout and intuitive interface. For improved usability, incorporate colour coding and visual clues.

Step 7: Use checkboxes for things like exercise routines and smoking history, dropdown menus for things like gender and chest pain kind, and textboxes for things like age and BMI to logically arrange the input fields.

Step 8: Using a binary output of 0/1, display the prediction findings and indicate whether the user is at high risk for diabetes and heart disease.

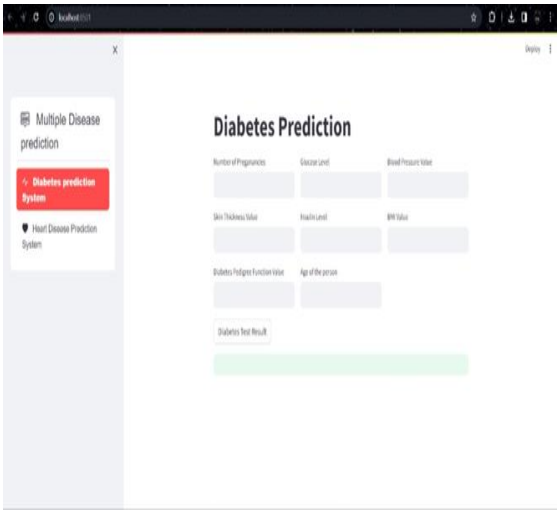

Fig.9 Diabetics portal

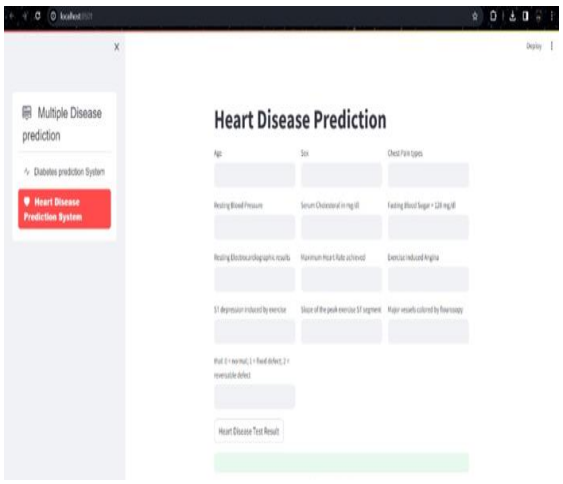

Fig.10 Heart disease portal

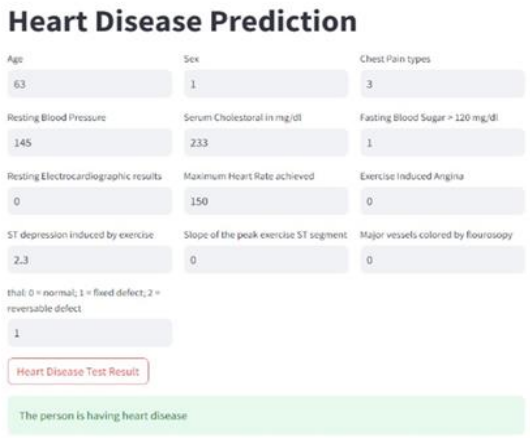

Fig.12 Output portal for heart disease

OUTPUT

Thus using the machine learning algorithms and following up the methodologies as said, we have successfully produced a system where we can accurately predict the diseases.

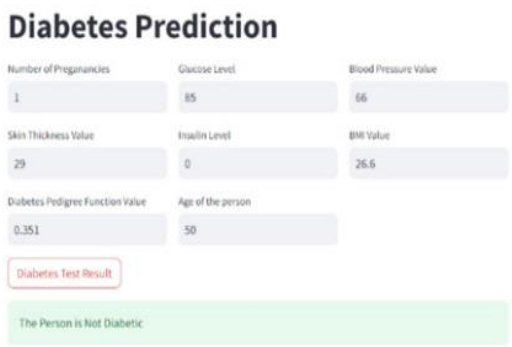

Fig.11 Output portal for diabetics

CONCLUSION

To sum up, the creation and implementation of the diabetes and heart disease illness prediction systems represent a major breakthrough in proactive healthcare management. The system provides a comprehensive tool for people to evaluate their health risks and make educated decisions by integrating predictive modelling, graphical user interfaces, and machine learning algorithms. A strong basis for collecting intricate correlations in health data is formed by the application of Logistic Regression and Support Vector Machines (SVM) methods. The graphical interface is easy to use and enhances user involvement by streamlining the input process, presenting clear visuals, and providing individualized forecasts. Important components of keeping usability include making sure that feedback mechanisms, results interpretable, and users may interact with the system. The system contributes to illness prevention and general well-

being by offering actionable advice to promote healthier lifestyles in addition to its educational role in revealing health risk factors. The illness prediction system, with its combination of data science, machine learning, and user-friendly interfaces, ultimately symbolizes the promise of technology in supporting preventative healthcare by enabling people to take proactive steps toward improved health.

REFRENCES

1. Bianca de Almeida-Pititto, Patrícia M. Dualib, Lenita Zajdenverg, JoanaRodrigues Dantas,Filipe Dias de Souza , Melanie Rodacki and Marcello Casaccia Bertoluci on behalf of Brazilian Diabetes Society Study Group (SBD), "Severity And Mortality Of COVID 19", In Patients With Diabetes, Hypertension And CardiovascularDisease: A Meta-analysis, Diabetology & Metabolic Syndrome Research, (2020)

2. R. Indrakumari, T. Poongodi, Soumya Ranjan Jena, "Heart Disease Prediction using Exploratory Data Analysis, International Conference" on Smart Sustainable Intelligent Computingand Applications under (ICITETM 2020)

3. Senthilkumar Mohan 1, Chandrasegar Thirumalai1, And Gautam Srivastava 2,3, (Member,IEEE), "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques",(IEEE ACCESS)

4. Aishwarya Majumdar, Dr. Vaidehi , "Diabetes Prediction using Machine Learning Algorithms, International Conference" On Recent Trends In Advanced Computing (2019, ICRTAC2019)

5. Emma Barron, Chirag Bakhai, Partha Kar, Andy Weaver, Dominique Bradley, Hassan Ismail, Peter Knighton, Naomi Holman, Kamlesh Khunti, Naveed Sattar, Nicholas J Wareham,Bob Young, Jonathan Valabhji, Associations of type 1 and type 2 diabetes with COVID-19 related mortality in England: a whole-population study, (Lancet Diabetes Endocrinol 2020)

6. Quan Zou, Kaiyang Qu , Yamei Luo, Dehui Yin, Ying Ju,Hua Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," Bioinformatics and Computational Biology, (Frontier Genetics Journal, 2018)

7. Md. Kamrul Hasan , Md. Ashraful Alam , Dola Das, Eklas Hossain, (Senior Member, IEEE), And Mahmudul Hasan, "Diabetes Prediction Using Ensembling of Different MachineLearning Classifiers", (IEEE ACCESS 2020)

8. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, Predictive Daa Mining for Medical Diagnosis", International Journal of Computer Applications, (Volume 17, #8,2011)

9. Himanshu Sharma, MA Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" ,International Journal on Recent and Innovation Trends in Computing and Communication (2016)

10. Baban U. Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare, Manisha Darade, "Heart Disease Prediction Using Machine Learning", IJARSCT, (2021)