IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

Disease Prediction System using Machine Learning Algorithms

Dr. Chhaya Yadav, Shreyash Tiwari, Nidhi Tiwari, Pulkit Agarwal, Abha Gupta
Department of Computer Science & Engineering
Raj Kumar Goel Institute of Technology, Ghaziabad, Uttar Pradesh, India

Abstract: *There are multiple techniques in machine learning that enable the analysis of large data sets from various industries. Predictive analytics in healthcare is hard work, but in the end, according to many documents, it can help practitioners make timely decisions about a patient's health and treatment. Diseases such as cancer. Diabetes and Heart disease kills many people worldwide, but these deaths are due to a lack of timely diagnosis. The lack of medical care in hospitals and the shortage of doctors to people cause the above problems. Diseases related to heart disease, cancer, and diabetes can pose a threat to humans if not detected early. This study included cancer, heart, and diabetes. To make this work uncomplicated and accessible to the masses, our team diagnosed websites using techniques from machine learning to make predictions for many types of diseases. Our goal in this study is to develop a web-based disease prediction application that uses machine learning-based prediction concepts to predict various diseases such as Diabetes, Parkinson's Disease, and heart disease. Our disease Prediction System is a Web Application used for predicting Human diseases by providing respective symptoms. Our system uses powerful machine learning algorithms to predict the disease based on the symptoms provided by the user [4].*

Keywords: Machine learning, SVM, Logistic Regression,Diabetes, Heart, Parkinson's, disease.

I. INTRODUCTION

In today's digital world, data is considered an asset, and the healthcare industry generates enormous amounts of data related to patient information. Many existing models focus on analyzing one disease per analysis, such as diabetes, cancer, or skin diseases, resulting in the need for separate systems to predict each disease. Thus, a common system capable of analyzing multiple diseases simultaneously would be highly beneficial. To address this need, we propose a system using Streamlit that can predict multiple diseases accurately and immediately based on the details provided by the user [3].

Disease prediction and diagnosis have undergone significant advancements in recent years due to the availability of large amounts of medical data and the development of advanced machine-learning algorithms. Machine learning algorithms

In today's digital world, data is considered an asset, and the healthcare industry generates enormous amounts of data related to patient information. Many existing models focus on analyzing one disease per analysis, such as diabetes, cancer, or skin diseases, resulting in the need for separate systems to predict each disease. Thus, a common system capable of analyzing multiple diseases simultaneously would be highly beneficial. To address this need, we propose a system using Streamlit that can predict multiple diseases accurately and immediately based on the details provided by the user [3].

Disease prediction and diagnosis have undergone significant advancements in recent years due to the availability of large amounts of medical data and the development of advanced machine-learning algorithms. Machine learning algorithms actions so more patients can get medicines within a shorter timeframe, thus saving the large number of lives [3].

Our system will initially focus on analyzing three diseases that are correlated with each other, namely Parkinson's disease, diabetes, and heart disease. By using machine learning algorithms and Streamlit, our system will enable users to input symptoms along with the disease name and obtain the corresponding diagnosis immediately. The user will access the API and send the disease parameters along with the disease name to invoke the corresponding model, which

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

will then return the patient's status. The main purpose of the proposed system is to accurately predict diseases based on the input provided by the user, using powerful machine learning algorithms. The project has real-time applications in the healthcare industry, enabling doctors and healthcare professionals to diagnose diseases quickly and accurately [1].

By using the Streamlit framework, the project provides an intuitive and user-friendly interface, making it accessible to a wide range of users. Overall, this project has the potential to revolutionize the healthcare industry by providing accurate and efficient disease prediction.

Our system will employ various machine learning algorithms, such as Naïve Bayes, and Support Vector Machine, to predict the accurate disease and determine which algorithm provides a faster and more efficient result. The primary benefit of our proposed system is that it includes all parameters that may cause the disease, making it more efficient and accurate in detecting the disease.

Finally, we will save the final model's behavior as a Python pickle file, which can be used for future predictions. In conclusion, our proposed system will provide immediate and accurate disease predictions to users, saving them the trouble of navigating various sites to predict different diseases.

II. LITERATURE REVIEW

TABLE I: COMPARISON OF VARIOUS METHODOLOGIES SUGGESTED BY AUTHORS

S. No.	Paper Name	Author	Year of Publish	Methodology
1.	Multiple Disease Prediction Model Using Machine Learning	D. Vasvi, D. Venkatesh, S. Santhosh Kumar, S. Sahaja, V.Santhosh Kumar	2023	Understand the designing of the model which gives accurate results
2.	Designing a Web application to predict disease	Samarth Dey, Mrs.. Priyanka Sonar & Anjali K Jaya Malini	2022	To analyze the different approaches to diagnosis the Diseases through web application.
3.	GDPS – General Disease Prediction System	Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E	2019	To understand the risk scores during using the various Algorithms.
4.	Disease prediction using Machine Learning over Big Data	Vinitha S, Sweetlin S, Vinusha H, Sajini S	2019	Advantages of using Machine learning algorithms over Big Data
5.	Disease Prediction. Using SVM Machine Learning Algorithm	Palle Pramod Reddy, Dr. Shivi Sharma, Hardeep kaur	2018	. SVM over CNN algorithm.
6.	Prediction of disease using Random Forest Classification	J. Senthil Kumar, S. Appavu	2018	To understand system as, “disease Recommendation system”

III. METHODOLOGY

Machine Learning:

Machine learning is a subfield of artificial intelligence that involves teaching computers to learn from data and make predictions or decisions based on that data, without being explicitly programmed. It involves building and training models that can analyze and extract patterns from data, and then using those models to make predictions or decisions about new data. Machine learning is used in a variety of applications, including image recognition, natural language processing, recommendation systems, and predictive analytics.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

SVM:

Support Vector Machine (SVM) is a powerful machine learning algorithm that is commonly used for classification and regression analysis in many fields, including image and text classification, bioinformatics, and finance. The algorithm has been widely adopted due to its ability to handle high-dimensional data and its robustness to noise. In SVM, the objective is to find the best boundary or hyperplane that separates different classes of data points. The algorithm achieves this by mapping the data into a high-dimensional feature space where it becomes easier to find a linear separation. Then, the algorithm selects the hyperplane that maximizes the margin, which is the distance between the hyperplane and the closest data points from both classes. The margin is important because it provides a good generalization for new data.

SVM is particularly useful when there is a need to balance the trade-off between model complexity and generalization ability. This is achieved by controlling the regularization parameter, which determines the level of importance given to the margin and the extent of slackness allowed for misclassifications. Furthermore, SVM allows for the use of different kernel functions, which can be used to transform the data into a higher dimensional space where a linear separation is possible. This allows SVM to handle non- linearly separable data [2].

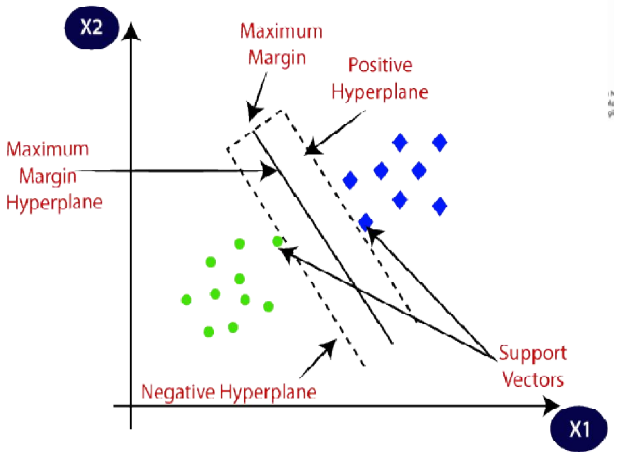

Fig.1. Support Vector Machine

Logistic regression:

Logistic regression is a highly popular algorithm in the field of machine learning and falls under the category of supervised learning techniques. Its primary purpose is to predict the outcome of a categorical dependent variable based on a given set of independent variables. The dependent variable should have discrete or categorical values such as Yes or No, 0 or 1, or true or false. However, instead of providing exact 0 or 1 values, logistic regression produces probabilistic values ranging between 0 and 1. While similar to linear regression, logistic regression is specifically suited for solving classification problems rather than regression problems.

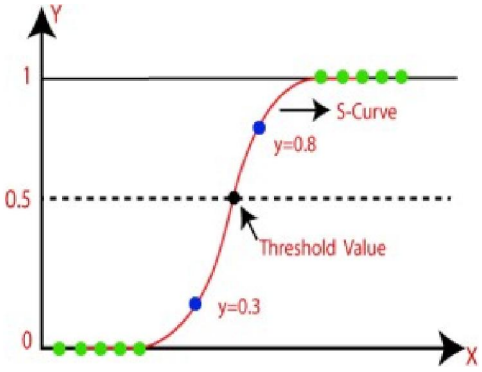

Fig.2. Logistic regression

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

In logistic regression, instead of fitting a traditional regression line, we employ an "S" shaped logistic function. This function estimates the likelihood of a specific outcome, such as determining whether cells are cancerous or not, or whether a mouse is obese based on its weight.

Logistic regression effectively classifies observations using various types of data and identifies the most influential variables for the classification task. The diagram below illustrates the logistic function [3].

About Dataset:

In a disease prediction system, the data preprocessing steps are crucial to ensure accurate and reliable predictions. For instance, cleaning the data can help to address issues such as missing or incorrect medical records, outliers, and duplicate patient data. Transforming the data can involve converting categorical patient data, such as symptoms and medical history, into numerical data that can be used for machine learning algorithms. Data integration can be helpful in providing a more comprehensive view of the patient's health status and risk factors, by combining data from different sources such as electronic health records andpatient surveys [5].

To put a dataset into a format that can be used by machine learning algorithms, the data needs to be converted into numerical values and arranged in a tabular format.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFu	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

Fig.3. Description of Diabetes Dataset

The dataset used in proposed system for diabetes disease prediction has attributes such as Pregnancy, glucose, blood pressure, skin thickness, insulin, BMI (body mass index), diabetes pedigree function, age.

The 9th attribute is class variable of each data points. Thisclass variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics [5].

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
49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
64	1	3	110	211	0	0	144	1	1.8	1	0	2	1

Fig.4. Description of Heart Disease Dataset

The dataset consists of 303 individuals' data. There are 14columns in the dataset, which has attributes such as age, sex, Chest-pain type, Resting blood pressure, Serum Cholesterol, Fasting blood sugar, Rest ECG, max heart rate. Exercise induced angina, Diagnosis of heart disease [8].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP
phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.0037
phon_R01_S01_2	122.4	148.65	113.819	0.00968	0.00008	0.00465
phon_R01_S01_3	116.682	131.111	111.555	0.0105	0.00009	0.00544
phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502
phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655
phon_R01_S01_6	120.552	131.162	113.787	0.00968	0.00008	0.00463
phon_R01_S02_1	120.267	137.244	114.82	0.00333	0.00003	0.00155
phon_R01_S02_2	107.332	113.84	104.315	0.0029	0.00003	0.00144
phon_R01_S02_3	95.73	132.068	91.754	0.00551	0.00006	0.00293
phon_R01_S02_4	95.056	120.103	91.226	0.00532	0.00006	0.00268
phon_R01_S02_5	88.333	112.24	84.072	0.00505	0.00006	0.00254
phon_R01_S02_6	91.904	115.871	86.292	0.0054	0.00006	0.00281
phon_R01_S04_1	136.926	159.866	131.276	0.00293	0.00002	0.00118
phon_R01_S04_2	139.173	179.139	76.556	0.0039	0.00003	0.00165
phon_R01_S04_3	152.845	163.305	75.836	0.00294	0.00002	0.00121
phon_R01_S04_4	142.167	217.455	83.159	0.00369	0.00003	0.00157
phon_R01_S04_5	144.188	349.259	82.764	0.00544	0.00004	0.00211
phon_R01_S04_6	168.778	232.181	75.603	0.00718	0.00004	0.00284
phon_R01_S05_1	153.046	175.829	68.623	0.00742	0.00005	0.00364

Fig.5. Description of Parkinson's Disease Dataset

The dataset contains a range of demographic, temporal, and biomedical voice measurements from 42 individuals recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring. The dataset includes 5,875 voice recordings from these individuals, providing a rich and diverse set of data for training and testing predictive models.

The primary objective of the dataset is to utilize the 16 voice measures to make predictions on the motor and total UPDRS scores. UPDRS (Unified Parkinson's Disease Rating Scale) is a widely used clinical rating scale for Parkinson's disease that measures motor and non-motor symptoms. Motor UPDRS specifically focuses on motor symptoms, such as tremors, rigidity, and bradykinesia, while total UPDRS encompasses both motor and non-motor symptoms [6].

The dataset used in proposed system for diabetes disease prediction has attributes such as age, sex, test time, motor_UPDRS, total_UPDRS, Jitter, Shimmer: APQ3, APQ5. APQ11, NHR, HNR, RPDE, DFA, PPE.

IV. ANALYSIS AND RESULT

In this section, we discussed the results of different ML algorithms applied to the dataset to predict a particular disease like SVM algorithm for Diabetes, the Logistic regression algorithm for heart disease, SVM algorithm Parkinson's disease.

In this study, we developed a disease prediction system using support vector machine (SVM) and logistic regression (LR) algorithms to predict a particular disease in patients. Diabetes disease prediction model in the system uses the Support vector machine algorithm with an accuracy score of 0.77, heart disease model uses the logistic regression algorithm with an accuracy score of 0.81 [16], and the Parkinson's disease model uses the Support vector machine algorithm with an accuracy score of 0.87. When the patient adds the parameter based on the disease parameters, it will show the range of values required as well as whether the patient has an illness or not based on the disease selected.

Overall, our findings show that a disease prediction system based on SVM and LR algorithms may reliably forecast the incidence of a certain disease in patients. The use of feature selection, evaluation on unbalanced data, and performance comparison of SVM and LR models might give doctors with useful insights for making better patient caredecisions.

Disease	Algorithm	Accuracy score
Diabetes	SVM	0.77
Heart Disease	Logistic Regression	0.81
Parkinson's Disease	SVM	0.87

Fig.6. Result of diseases using various Algorithms

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

V. CONCLUSION

In conclusion, Early disease diagnosis and treatment are facilitated by machine learning algorithms, improving patient outcomes. This research reviewed existing literature on disease prediction using these algorithms, focusing on their advantages and disadvantages. Our unique approach incorporates feature selection, model interpretability, ensemble learning, and addressing imbalanced data, distinguishing it from previous methods. These techniques enhance the accuracy and comprehensibility of disease predictions, assisting clinicians in making informed decisions. Future research can explore the application of explainable AI and causal inference techniques to further improve disease prediction. Additionally, our system can be transformed into a mobile app for patients, enabling them to input symptoms and receive potential disease predictions. Integration with other diagnostic tools, like imaging and laboratory tests, would provide a more comprehensive diagnostic approach [9].

In the future, our system can be transformed into a mobile app for patients to input symptoms and receive disease predictions. Integration with imaging and laboratory tests will enhance the diagnostic process, offering a comprehensive approach

REFERENCES

[1] Kumari, V. Anuja, and R. Chitra. "Classification of diabetes disease using support vector machine." International Journal of Engineering Research and Applications 3.2 (2013): 1797-1801.

[2] Sisodia, Deepti, and Dilip Singh Sisodia, "Prediction of diabetes using classification algorithms." Procedia computer science 132 (2018): 1578-1585.

[3] Tafa, Zhilbert, Nerxhivane Pervetica, and Bertran Karahoda. "An Conference on Embedded Computing (MECO). IEEE, 2015.

[4] Mujumdar, Aishwarya, and V. Vaidehi."Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292299.

[5] Joshi, Tejas N., and P. P. M. Chawan."Diabetes prediction using machine learning techniques." Ijera 8.1 (2018): 9-13.

[6] Sriram, T. V., et al. "Intelligent Parkinson disease prediction using machine learning algorithms." Int. J. Eng. Innov. Technol 3 (2013): 212-215.

[7] Yadav, Anupama, Levish Gediya, and Adnanuddin Kazi. "Heart disease prediction using machine learning." International Research Journal of Engineering and Technology (IRJET 8.09 (2021).

[8] Jindal, Harshit, et al. "heart disease prediction using machine learning algorithms." IOP conference series: materials science and engineering. Vol. 1022. No. 1. IOP Publishing, 2021.

[9] Sharma, Vijeta, Shrinkhala Yadav, and Manjari Gupta. "Heart disease prediction using machine learning techniques." 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). IEEE, 2020.

[10] Arumugam, K., et al. "Multiple disease prediction using Machine

[11] Priyanka Sonar, Prof. K. Jaya Malini," DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES", 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication (ICCMC) .

[12] Achana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3).

[13] Mir, S.N. Dhage, in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (IEEE, 2018).

[14] Y. Khourdi, M. Bahaj, Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization, Int. J. Intell. Eng. Syst. 12(1), 242 (2019).

[15] S. Vijayarani, S. Dhayanand, Liver disease prediction using SVM and Naive bayes algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) 4(4), 816 (2015)

[16] S. Mohan, C. Thirumalai, G. Srivastava, Elective heart disease prediction using hybrid machine learning techniques, IEEE Access 7, 81542 (2019)

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 11, May 2023

[17] T.V. Sriram, M.V. Rao, G.S. Narayana, D. Kaladhar, T.P.R. Vital, Intelligent Parkinson disease prediction using machine learning algorithms, International Journal of Engineering and Innovative Technology (IJEIT) 3(3),1568 (2013)
[18] A.S. Monto, S. Gravenstein, M. Elliott, M. Collopy, J. Schweinle, Clinical signs and symptoms predicting influenza infection, Archives of
[19] R.D.H.D.P. Sreevalli, K.P.M. Asia, Prediction of diseases using random forest classification algorithm
[20] D.R. Langbehn, R.R. Brinkman, D. Falush, J.S. Paulsen, M. Hayden, an International Huntington's Disease Collaborative Group, A new model for prediction of the age of onset and penetrance for Huntington's disease based on cag length, Clinical genetics 65(4), 267 (2004)