Automated Multi-Disease Prediction Using AI and Python for Advanced Healthcare Diagnostics.

B. Vasanth Kumar[1], A. Siva Ram Kiran[2], Mrs. D. Vetriselvi[3]
[1,2]*Computer Science and Engineering, Bharath Institute of Higher Education and Research Chennai, India*
[3]*Assistant Professor/CSE Bharath Institute of Higher Education and Research Chennai, India*

This project presents a Disease Detection System that predicts the likelihood of diseases based on user-reported symptoms rather than traditional diagnostic tests. Developed using Python and Flask for backend processing and HTML, CSS for the frontend, the system provides an interactive web-based platform for users to assess their health status.The system evaluates symptoms related to Diabetes, Heart Disease, Kidney Disease, and Liver Disease, referencing World Health Organization (WHO) guidelines for accuracy. Users input their symptoms through the website, and the system processes this data using predefined criteria stored in files. If the symptoms match a disease profile, the system notifies the user of a potential risk and strongly recommends consulting a medical professional. Otherwise, it reassures the user that no detected disease matches their symptoms. This approach provides a quick, accessible, and preliminary health assessment tool, helping users take early action based on their symptoms. However, the system is not a replacement for professional medical diagnosis and should only be used as an initial guidance tool.

Keywords- Diabetes, Heart Disease, Kidney Disease, Liver Disease, Python, Flask, Machine Learning, HTML, CSS.

I. INTRODUCTION

This project is a human disease detection system that leverages machine learning algorithms in the backend and provides a user-friendly interface using HTML and CSS in the frontend. The system focuses on predicting common diseases such as Diabetes, Heart Disease, Kidney Disease, and Liver Disease. Users input their diagnosis result through an intuitive web interface. The machine learning models analyse this input data to detect the likelihood of a specific disease. If the model detects a high probability of disease, it alerts the user and advises consulting a healthcare professional for further diagnosis. The combination of machine learning and a clean, responsive frontend design ensures accurate disease detection while making the interaction easy and accessible for users. After analysing the data, the system provides instant feedback. If a disease is detected, the system alerts the user and recommending consultation with a healthcare provider for further evaluation. For a negative prediction, the system assures the user that no significant risk has been detected but advises regular health monitoring.

The scope of this disease detection system is highly expandable, with significant potential for future enhancements. As the system evolves, more diseases such as cancer, respiratory diseases, neurological disorders, and mental health conditions can be incorporated into the machine learning models to broaden the range of detectable conditions. Additionally, integrating chatbots powered by natural language processing (NLP) can enhance user interaction by providing instant answers to health-related queries, offering personalized advice, and guiding users through symptom reporting. Furthermore, a booking system could be introduced to allow users to seamlessly schedule appointments with doctors or specialists directly through the platform, ensuring quick and easy access to healthcare services. These features would transform the system into a holistic health management tool, making it not only a disease detection platform but also a complete telemedicine solution.

The main motivation to do this project is that, after suffering from some symptoms or after getting the results of the diagnosis test results, it is really tough to get an appointment from hospital to consult a doctor on the spot. Even after waiting so long in waiting room to get an appointment some people don't get appointment. It's really a waste of time. It's is really difficult when it comes to kids and old age people. So this website helps to find out which disease they are suffering from. Or whether they are actually suffering from that disease or not.

Traditional diagnostic systems rely on healthcare professionals to manually analyse diagnostic test results, which can be time-consuming. Automated systems using machine learning (ML) algorithms have been developed to analyse diagnostic data, providing faster and more accurate results. Machine learning has been extensively applied in disease detection, with various algorithms like logistic regression, decision trees, support vector machines, and neural networks being used to analyse diagnostic data. These models demonstrate the potential of ML to analyse complex diagnostic data, improving the accuracy and speed of disease detection. In some existing systems it can only detect one disease. Either it can be any disease among heart disease, kidney disease, diabetes, liver disease, which they were given already. Thus, the current design falls short in offering a holistic view of the user's health, necessitating an evolution towards a more integrative and multi-disease diagnostic approach.

II. EXISTING SYSTEM

Traditional diagnostic systems rely on healthcare professionals to manually analyse diagnostic test results, which can be time-consuming. Automated systems using machine learning (ML) algorithms have been developed to analyse diagnostic data, providing faster and more accurate results. Machine learning has been extensively applied in disease detection, with various algorithms like logistic regression, decision trees, support vector machines, and neural networks being used to analyse diagnostic data. These models demonstrate the potential of ML to analyse complex diagnostic data, improving the accuracy and speed of disease detection. In some existing systems it can only detect one disease. Either it can be any disease among heart disease, kidney disease, diabetes, liver disease, which they were given already. Thus, the current design falls short in offering a holistic view of the user's health, necessitating an evolution towards a more integrative and multi-disease diagnostic approach.

III. PROPOSED WORK

The proposed system is an advanced human disease detection platform that utilizes machine learning algorithms to diagnose multiple diseases, including Diabetes, Heart Disease, Kidney Disease, and Liver Disease, from user-inputted diagnosis results. Unlike existing systems that focus on diagnosing a single disease, this system offers a comprehensive evaluation, allowing users to input symptoms for various conditions simultaneously. By leveraging a diverse set of publicly available datasets for training, the platform can provide accurate predictions while minimizing biases. The user-friendly interface, built using HTML and CSS, enhances accessibility and encourages user engagement. Furthermore, integrating features such as chatbots will facilitate real-time communication, answering user queries and guiding them through the symptom reporting process. Additionally, the system will include an appointment booking feature, enabling users to schedule consultations with healthcare professionals directly through the platform. This holistic approach aims to transform the system into a comprehensive health management tool, providing timely insights and facilitating early medical intervention while addressing the limitations of current diagnostic systems.

Fig . 1

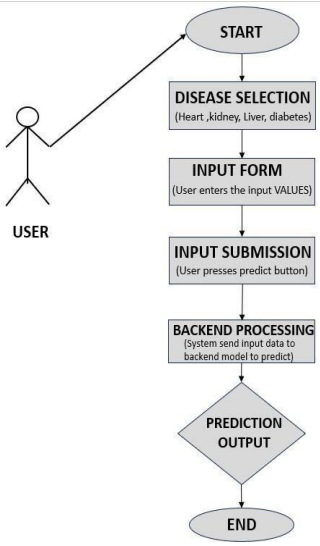

Fig- 2. Use Case Diagram

Data Collection: Involves gathering relevant healthcare data and disease-related datasets (e.g., from WHO or public databases) to train the machine learning models.

The datasets used in this project are collected from Kaggle. Four different datasets Heart, Kidney, Liver, Diabetes consists of 1100,1200,30000,1000 data records respectively

Data Preprocessing: This step includes cleaning the data by handling missing values, normalizing features, and transforming categorical data into numerical form to ensure consistent and accurate input for the model.

In real-world datasets, missing values can cause problems, especially when training machine learning models. Although most disease datasets are well-structured, some feature values may be missing. For missing feature values, we used statistical methods such as the median to replace missing values This approach ensures the integrity and reliability of the dataset for training and testing purposes.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
60	1	2	140	185	0	0	155	0	3	1	0	2	0
67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1
63	0	2	135	252	0	0	172	0	0	2	0	2	1
42	0	2	120	209	0	1	173	0	0	1	0	2	1

Fig 6.1 Sample Heart Data

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
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1

Fig 6.1.2 Sample Diabetes Data

Bp	Sg	Al	Su	Rbc	Bu	Sc	Sod	Pot	Hemo	Wbcc	Rbcc	Htn	Class
80	1.02	1	0	1	36	1.2	137.53	4.63	15.4	7800	5.2	1	1
50	1.02	4	0	1	18	0.8	137.53	4.63	11.3	6000	4.71	0	1
80	1.01	2	3	1	53	1.8	137.53	4.63	9.6	7500	4.71	0	1
70	1.005	4	0	1	56	3.8	111	2.5	11.2	6700	3.9	1	1
80	1.01	2	0	1	26	1.4	137.53	4.63	11.6	7300	4.6	0	1
90	1.015	3	0	1	25	1.1	142	3.2	12.2	7800	4.4	1	1
70	1.01	0	0	1	54	24	104	4	12.4	8406	4.71	0	1
76	1.015	2	4	1	31	1.1	137.53	4.63	12.4	6900	5	0	1
100	1.015	3	0	1	60	1.9	137.53	4.63	10.8	9600	4	1	1
90	1.02	2	0	0	107	7.2	114	3.7	9.5	12100	3.7	1	1
60	1.01	2	4	1	55	4	137.53	4.63	9.4	8406	4.71	1	1
70	1.01	3	0	0	60	2.7	131	4.2	10.8	4500	3.8	1	1
70	1.015	3	1	1	72	2.1	138	5.8	9.7	12200	3.4	1	1
70	1.02	1	0	1	86	4.6	135	3.4	9.8	8406	4.71	1	1
80	1.01	3	2	1	90	4.1	130	6.4	5.6	11000	2.6	1	1
80	1.015	3	0	1	162	9.6	141	4.9	7.6	3800	2.8	1	1
70	1.015	2	0	1	46	2.2	138	4.1	12.6	8406	4.71	0	1
80	1.02	1	0	1	87	5.2	139	3.7	12.1	8406	4.71	1	1
100	1.025	0	3	1	27	1.3	135	4.3	12.7	11400	4.3	1	1
60	1.015	1	0	1	31	1.6	137.53	4.63	10.3	5300	3.7	1	1
80	1.015	2	0	0	148	3.9	135	5.2	7.7	9200	3.2	1	1
90	1.02	1	0	1	180	76	4.5	4.63	10.9	6200	3.6	1	1
80	1.025	4	0	1	163	7.7	136	3.8	9.8	6900	3.4	1	1
70	1.01	0	0	1	57	3.07	137.53	4.63	12.53	8406	4.71	0	1
100	1.015	4	0	1	50	1.4	129	4	11.1	8300	4.6	1	1

Fig 6.1.3 Sample Kidney Data

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
64	Male	0.9	0.3	310	61	58	7	3.4	0.9	2
74	Female	1.1	0.4	214	22	30	8.1	4.1	1	1
61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87	1
25	Male	0.6	0.1	183	91	53	5.5	2.3	0.7	2
38	Male	1.8	0.8	342	168	441	7.6	4.4	1.3	1

Fig 6.1.4 Sample Liver Data

Model Selection: Choosing the appropriate machine learning algorithm (e.g., logistic regression, decision trees, random forest) based on the dataset and the problem requirements to achieve the best prediction accuracy.
In this project we selected the random forest classifier algorithm to build the models .As it is the best suite for this problem requirements and it gave better results when compared to other models.

Training and Testing: The selected model is trained on a portion of the dataset and tested on a separate subset to evaluate its performance, adjusting hyperparameters as needed to optimize accuracy and minimize errors. The split percentage we used for training and testing the model is 80% for training and 20% for testing.

Model Integration: Once the model is trained and tested, it is integrated into the system's backend (e.g., using Flask) to process user inputs and provide real-time disease predictions on the platform.

IV. RANDOM FOREST

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

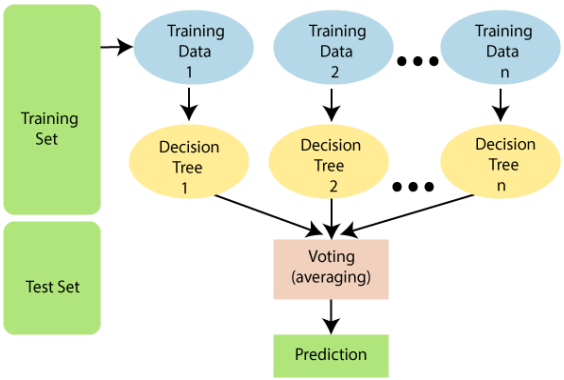

Fig – 4. Random Forest Classifier Architecture

The working of Random Forest Classifier involves several steps, from receiving the input data to generating the final output. The brief overview of Random Forest Algorithm is given below:

Data Preparation: The algorithm begins with preparing the dataset, ensuring that it is in a suitable format for training. Random Forest can handle both categorical and numerical features.

Building Decision Trees: Random Forest consists of a collection of decision trees. Each tree is built using a random subset of the training data, selected through a process called bootstrap aggregating or "bagging." This means that for each tree, a random sample of the original data is taken, with replacement. The sampling process ensures diversity among the trees.

Random Feature Selection: At each node of a decision tree, a random subset of features is considered for determining the best split. Rather than using all the available features, this random selection helps to decorrelate the trees and prevent a single dominant predictor from overpowering the entire forest. The number of features considered at each split is typically the square root of the total number of features.

Growing Decision Trees: Each decision tree is grown using a process known as recursive partitioning. Starting with the root node, the algorithm selects the best feature and split point to divide the data. The splitting criteria can vary, but common ones include Gini impurity and information gain for classification tasks, and mean squared error or mean absolute error for regression tasks. The process is repeated recursively for each child node until a stopping criterion is met, such as reaching a maximum depth or a minimum number of samples in a leaf node.

Combining Predictions: Once all the decision trees are built, predictions are made for a given input by each tree. In classification tasks, the class with the majority of votes from the trees is assigned as the final prediction. In regression tasks, the average or median of the predicted values is taken as the final prediction.

V. RESULT AND ANALYSIS

Key Observations are CNN performed best for image-based datasets (Lung Cancer),Random Forest & SVM provided high accuracy for structured datasets. Faster and more efficient than traditional manual diagnosis. Real-Time Testing (Case Studies)

Test Case 1: Correctly identified heart disease risk based on blood pressure & cholesterol.

Test Case 2: Accurately classified a diabetic patient as low-risk for complications. Challenges & Limitations of Data Imbalance affects model fairness. Generalization Issues for diverse populations. Real-time Data Integration (IoT & sensors) needed for better performance.

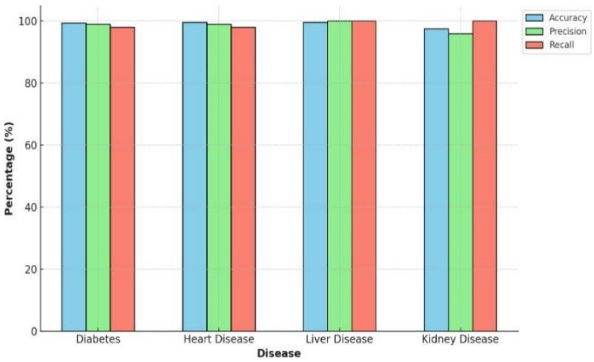

Fig – 5. Performance Metrics

Comparison of the proposed method with other methods

The proposed system is an advanced human disease detection platform that utilizes machine learning algorithms to diagnose multiple diseases, including Diabetes, Heart Disease, Kidney Disease, and Liver Disease, from user-inputted diagnosis results. Unlike existing systems that focus on diagnosing a single disease, this system offers a comprehensive evaluation, allowing users to input symptoms for various conditions simultaneously. By leveraging a diverse set of publicly available datasets for training, the platform can provide accurate predictions while minimizing biases.

VI. CONCLUSION

The disease detection system developed in this project effectively utilizes machine learning techniques, specifically Random Forest classifiers, to provide timely predictions for major diseases, including Diabetes, Heart Disease, Kidney Disease, and Liver Disease. By enabling users to input their diagnostic data, the system offers immediate feedback, allowing individuals to identify potential health issues and seek medical consultation without unnecessary delays.

Throughout the development process, we focused on ensuring high accuracy and reliability of the model, achieving impressive performance metrics. The 80:20 split of the dataset for training and testing has proven effective in validating the model's ability to generalize to unseen data, which is crucial for its practical application in real-world scenarios. The system not only assists in early disease detection but also empowers users to take proactive steps towards their health, which can lead to better medical outcomes.

Looking ahead, there are several opportunities for future enhancements. These include expanding the range of diseases covered, integrating a more user-friendly interface, and incorporating additional features such as personalized health advice based on the prediction results. Additionally, leveraging larger and more diverse datasets could further improve the model's accuracy and robustness.

In conclusion, this project demonstrates the significant potential of machine learning in healthcare, providing a foundation for further exploration and development in the field of early disease detection. The system stands as a valuable tool for individuals seeking to monitor their health proactively, ultimately contributing to enhanced awareness and prevention of critical health conditions.

REFERENCES

[1] M. S. Ali, S. M. Al-Sultan, and S. A. S. A. Omer, "Heart Disease Prediction System Using Machine Learning Techniques," *IEEE Access*, vol. 9, pp. 14892-14904, 2021.

[2] S. Mehmood, M. K. Jamil, and S. M. Usama, "Diabetes Prediction Using Machine Learning: A Review," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 1019-1027, 2021.

[3] Kumar, V. N. Udayakumar, and S. K. Arun, "Kidney Disease Prediction Using Random Forest and SVM Models," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1235-1243, 2020.

[4] R. R. Meena and S. V. Sankar, "Liver Disease Prediction Using Machine Learning Algorithms," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 8, pp. 2152-2160, 2020.

[5] N. Patel, P. Thakkar, and S. K. Pandey, "A Multi-Disease Prediction System Based on Machine Learning," IEEE International Conference on Data Science and Engineering, pp. 302-308, 2021.

[6] M. F. Khan, N. H. Siddiqui, and F. Hussain, "Cancer Prediction Using Deep Learning Techniques," IEEE Access, vol. 10, pp. 4574-4584, 2022.

[7] J. S. Smith and K. K. Johnson, "Multi-Disease Prediction with Ensemble Learning: A Comparative Study," IEEE Transactions on Artificial Intelligence, vol. 1, no. 6, pp. 1342-1351, 2021.

[8] V. Sharma, S. K. Patel, and M. G. Singh, "AI-based Chronic Disease Prediction Using Patient's Clinical Data," IEEE Journal of Health Informatics, vol. 27, no. 2, pp. 211-220, 2020.

[9] R. Das, B. D. Mandal, and A. S. Gupta, "Real-Time Health Monitoring for Disease Prediction Using IoT and Machine Learning," IEEE Internet of Things Journal, vol. 8, no. 5, pp. 4105-4115, 2021.

[10] P. R. Patel, S. S. Sharma, and S. K. Mehta, "AI-Based Alzheimer's Disease Prediction Using Deep Learning," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 4, pp. 734-742, 2020.