SMART HEALTH: A MACHINE LEARNING APPROACH TO PREDICTING AND MANAGING EARLY LIFE STYLE DISEASES

PITCHUKA BHAVYA Department of CSE (Data Science), Sreyas Institute of Engineering and Technology Nagole, Hyderabad, Telangana 500068
KOTIH KAMAKSHI Department of CSE (Data Science), Sreyas Institute of Engineering and Technology Nagole, Hyderabad, Telangana 500068
MARIKANTI NIKITHA Department of CSE (Data Science), Sreyas Institute of Engineering and Technology Nagole, Hyderabad, Telangana 500068
BANOTH NITHIN Department of CSE (Data Science), Sreyas Institute of Engineering and Technology Nagole, Hyderabad, Telangana 500068
Dr. G. NAGA RAMA DEVI Professor, CSE-DS, Sreyas Institute of Engineering and Technology, Nagole, Hyderabad, Telangana 500068

ABSTRACT

Our innovative web application harnesses the power of machine learning to revolutionize disease prediction, particularly targeting early lifestyle diseases such as obesity and type 2 diabetes, which are often linked to unhealthy habits. By employing advanced algorithms including Decision Trees, Logistic Regression, Naïve Bayes, and boosting techniques like Gradient Boosting and XGBoost, our system accurately predicts potential health issues based on user symptoms. The integration of Natural Language Processing (NLP) and an AI-powered chatbot ensures a seamless, user-friendly experience, facilitating online doctor appointments, access to prediction results, and easy medicine purchases. Furthermore, the application utilizes data logs to identify virus-affected areas, providing timely alerts to help users maintain safe distances and reduce the spread of illness. This comprehensive approach not only mitigates inefficiencies and security vulnerabilities of current prediction methods but also minimizes the need for hospital visits, enhancing healthcare accessibility and promoting proactive health management.

KEYWORDS

Lifestyle diseases, Machine learning, NLP, AI, Decision tree classifier

INTRODUCTION

Machine learning is a field of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to perform tasks without explicit instructions. Instead, these algorithms learn and improve from experience or data. It encompasses various techniques such as supervised learning, unsupervised learning, and reinforcement learning, allowing computers to recognize patterns, make predictions, and generate insights from data. Early lifestyle diseases resulting from poor habits like a sedentary lifestyle, unhealthy diet, and stress, leading to conditions such as obesity, diabetes, and cardiovascular issues.

Developed using the Flask framework and Python programming language with a MySQL database, our web application provides three main medical services: disease prediction, appointment booking, and access to medical information. By leveraging the capabilities of machine learning algorithms, our application facilitates precise disease prediction through two variants: general forecasting and specific disease prediction.

This empowers users to anticipate potential illnesses without immediate reliance on a doctor's consultation, thereby saving both time and healthcare costs. At the heart of our application lies a chatbot interface, which employs Natural Language Processing (NLP) techniques to promptly and efficiently address user queries. This interactive feature serves as a vital resource for individuals seeking disease-related information, enhancing accessibility to healthcare knowledge. Moreover, the platform streamlines the process of scheduling appointments with available doctors, promoting seamless access to medical assistance. Our overarching goal is to provide a comprehensive solution that enables users to monitor their health, obtain medical guidance, and schedule appointments through a single, user-friendly interface. By integrating machine learning technology with patient-centric features, our application endeavors to advance healthcare accessibility, facilitate early disease detection, and ultimately contribute to the prevention of chronic illnesses.

Furthermore, the incorporation of Natural Language Processing (NLP) and an AI-powered chatbot enhances the user experience by enabling seamless interaction and addressing disease-related queries in real-time. Users can access predicted health outcomes, receive guidance on managing symptoms, and even schedule online doctor appointments or purchase necessary medications through the platform. This holistic approach not only empowers individuals to take charge of their health but also streamlines the healthcare process, reducing the burden on traditional healthcare infrastructure.

This proactive approach to disease management not only promotes public health but also enhances the overall efficiency and accessibility of healthcare services by minimizing the need for unnecessary hospital visits.

I. LITERATURE SURVEY

The literature review presents a diverse array of chatbot applications in the healthcare sector, emphasizing their potential to enhance medical diagnostics, patient interaction, and mental health counseling. The paper on "An AI-Based Medical Chatbot Model for Infectious Disease Prediction" showcases a chatbot leveraging a deep feedforward multilayer perceptron with an impressive accuracy of 94.32%, utilizing TensorFlow to build its NLP and deep neural network architecture. This model highlights the efficacy of chatbots in predicting infectious diseases. Another study, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities" employs a CNN algorithm, achieving a 70% prediction accuracy. However, it notes regional disease characteristics as a limiting factor. Additionally, the "Chatbot for Psychiatric Counseling in Mental Healthcare Service" employs high-level NLU and a multi-modal approach for emotion recognition, achieving a 91.3% accuracy, although it lacks continuous user monitoring and ethical judgment in interventions.

Further, "Ratatta: Chatbot Application Using Expert System" focuses on keyword scanning for replies, utilizing an XML-based dataset from Stack Exchange and the K-Means clustering algorithm, achieving a 92% accuracy but facing efficiency issues in producing quick results. Lastly, "A Tool of Conversation: Chatbot" discusses the design and implementation of chatbot systems, highlighting their potential in solving health-related problems and predicting a 93% accuracy. This paper underscores the versatility and future dominance of chatbots, emphasizing their capability to offer flexible solutions and enhance AI utility in various applications. Overall, the literature illustrates the significant advancements in chatbot technology for healthcare, their varying methodologies, and respective strengths and limitations.

Chatbots can provide a new and flexible way for users. They are giving AI something better to do.

We will also study another application where Chatbots could be useful and techniques used while designing a Chatbot.

II. RELATED WORK

3.1PROPOSED APPROACH

Our proposed solution utilizes a lifestyle disease dataset sourced from Kaggle, where symptoms act as features and disease types as labels. Employing machine learning algorithms like Decision Trees (DT), Logistic Regression (LR), Naive Bayes (NB), and Support Vector Machines (SVM), along with boosting techniques such as gradient boosting and xgboosting via an ensemble method, we aim to train the dataset and compare their accuracies to identify the most efficient model for disease prediction. We have achieved best algorithm as DECISION TREE algorithm with 97% of accuracy among all other ml algorithms.

This chosen algorithm is integrated into a Flask web application, simulating a hospital scenario, enabling users to input symptoms and receive predictions for early lifestyle diseases. With the increasing prevalence of lifestyle diseases, our objective is to develop effective machine learning models to predict potential issues, thereby reducing the need for hospital visits. Enhancements include features for scheduling doctor appointments, purchasing medications online, and incorporating an AI chatbot to address health-related queries, thus minimizing the necessity of hospital visits. Additionally, we aim to leverage data to identify virus-affected areas, facilitating timely warnings for individuals to maintain safe distances. Our comprehensive online platform utilizes machine learning methods to forecast potential lifestyle illnesses, providing a user-friendly interface integrating natural language processing (NLP) and ensemble techniques, allowing users to schedule appointments, monitor well-being, and access medical information efficiently.

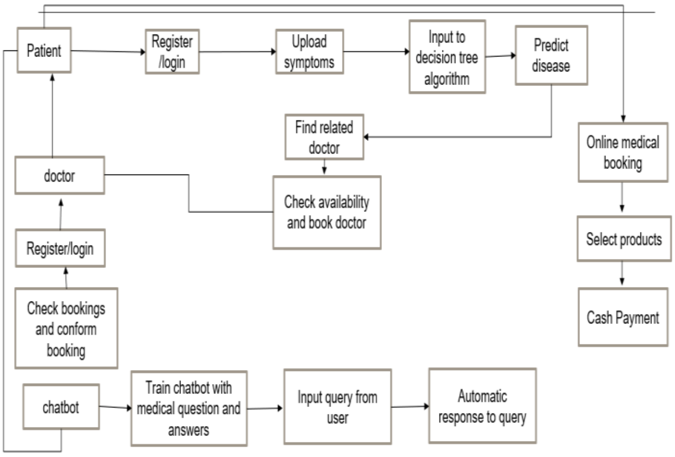

Patient
Register /login
Upload symptoms
Input to decision tree algorithm
Predict disease
Find related doctor
Online medical booking
doctor
Check availability and book doctor
Select products
Register/login
Check bookings and conform booking
Cash Payment
chatbot
Train chatbot with medical question and answers
Input query from user
Automatic response to query

Fig 1: Architecture of Predicting life style diseases using classifier techniques

3.2 ALGORITHMS

1. **DECISION TREE:** Decision Tree is a hierarchical structure used in machine learning and data mining for classification and regression tasks. It resembles a flowchart where each internal node represents a decision based on a feature attribute, each branch represents the outcome of the decision, and each leaf node represents the final decision or classification. Decision trees are popular due to their simplicity in visualization and interpretation, making them useful for both understanding complex decision-making processes and building predictive models

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

2. **NAIIVE BAYES:** Naiive Bayes is a simple probabilistic classifier based on Bayes' theorem with an assumption of independence between features. It calculates the probability of a given instance belonging to a particular class by multiplying the probabilities of each feature occurring given that class. Despite its naive assumption, Naive Bayes often performs well in practice, particularly with text classification tasks, due to its simplicity, speed, and ability to handle high-dimensional data efficiently. In the context of detection, prediction can help identify critical diseases that may indicate mental health concerns, allowing for more precise and context-aware predictions.

classifier

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

3. **LOGISTIC REGRESSION (LR):** Logistic Regression is a linear model used for binary classification tasks. It models the probability that an instance belongs to a particular class using the logistic function. Despite its simplicity, Logistic Regression is widely used for its interpretability and efficiency. Logistic Regression serves as a baseline model for binary classification, providing a straightforward and interpretable approach. It allows for an initial understanding of the classification

problem and can serve as a benchmark for evaluating the performance of more complex models.

$$y = \frac{e^{(b_0 + b_1X)}}{1 + e^{(b_0 + b_1X)}}$$

4. **XGBOOSTING:** XGBoosting (Extreme Gradient Boosting) is utilized as a boosting algorithm to enhance the predictive accuracy of the model. XGBoosting works by sequentially adding decision trees to correct the errors of the previous trees, thereby improving overall model performance. It effectively combines the strengths of decision trees with regularization techniques to prevent overfitting and optimize predictive power. By iteratively refining the model based on the residuals of previous iterations, XGBoosting achieves higher accuracy compared to standalone decision trees, making it a valuable component in disease detection and prediction tasks.

$$\text{Similarity Score} = \frac{\left(\sum \text{Residuals}\right)^2}{\sum^{N}[P\ (1 - P)] + \lambda}$$

5. **GRADIENT BOOSTING:** gradient boosting is used to enhance the predictive accuracy of the model. Initially, a weak predictive model is trained on the dataset. Subsequent models are then built to correct the errors of the previous ones, with each subsequent model focusing on the instances that were misclassified by the previous models.

This iterative process continues until a strong predictive model is created, resulting in improved performance for disease detection compared to using a single model alone.

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1,\ldots,n$$

In our project, we utilized a diseases dataset from Kaggle, achieving a remarkable 97% accuracy with a decision tree algorithm. The dataset categorizes diseases with binary labels, 0 and 1, indicating the presence or absence of a condition. Our comprehensive analysis involved both training and testing phases, using training.csv and testing.csv files, respectively. The decision tree model's exceptional performance underscores its efficacy in classifying diseases within this dataset.

It can significantly benefit patients by providing quicker and more accurate disease diagnoses, enabling timely and effective treatments. The decision tree model can assist healthcare professionals in identifying conditions with high precision, reducing the likelihood of misdiagnosis. It can also facilitate remote medical consultations by analyzing patient data and offering diagnostic insights. Additionally, the model's ability to process and interpret large volumes of data can aid in personalized medicine, treatments to individual patient needs and improving overall health outcomes.

By aiding healthcare professionals in identifying conditions with high precision and supporting remote medical consultations, the decision tree model can reduce the likelihood of misdiagnosis. Furthermore, its ability to handle and interpret large volumes of data paves the way for personalized medicine, ultimately improving overall health outcomes and patient care.

Data preprocessing involves cleaning the dataset by handling missing values and outliers, and performing feature scaling and normalization to ensure uniformity and improve model performance. The categorical variables are encoded into numerical format for compatibility with machine learning algorithms.

It utilizes machine learning techniques to predict and manage early lifestyle diseases effectively. By analyzing various lifestyle factors and health data, the system aims to provide early detection of diseases such as diabetes, hypertension, and obesity. The implementation involves developing predictive models that can identify individuals at risk of developing lifestyle-related diseases. Through proactive monitoring and personalized interventions, the project seeks to empower individuals to adopt healthier lifestyles and mitigate the risk of chronic diseases. Ultimately, the goal is to improve overall health outcomes and reduce the burden of lifestyle-related illnesses on healthcare systems.

III. RESULT AND DISCUSSION

Result

Using this system and approach, we successfully generated multiple services based on user-given symptom inputs. In our web application, we identified Decision Tree as the best algorithm, achieving 97% accuracy.

Algorithms	Accuracy Test	F1 Score	Recall
Decision Tree classifier Random Forest	97.51454	0.9195	0.8056
Classifier	95.12195	0.9394	0.8333
Logistic Regression	95.12195	0.8727	0.8050
Gradient Boosting	94.06	0.9311	0.8403
Xgboosting	94.06	0.9409	0.8467

Fig.1 Accuracy of algorithms

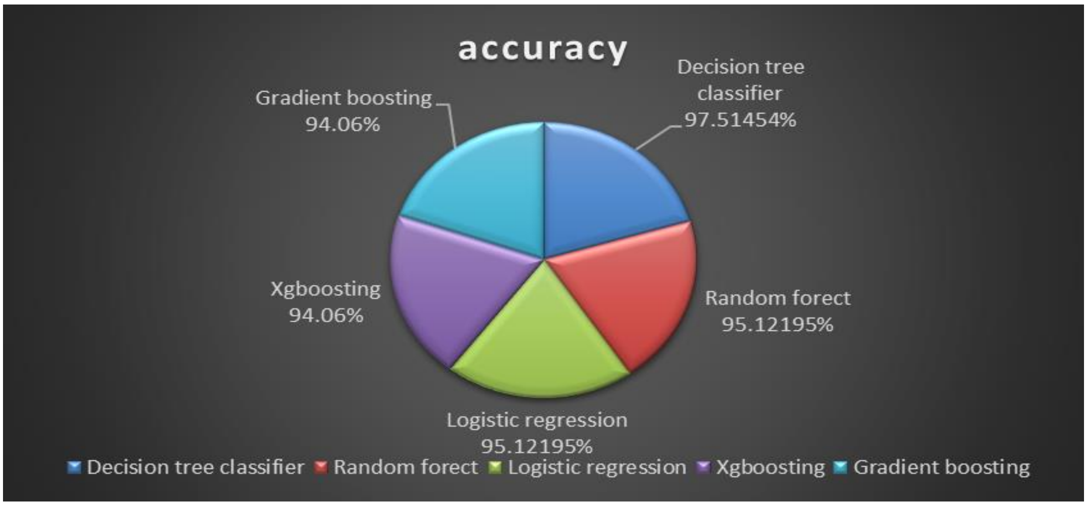


Fig.2 Piechart of Accuracy

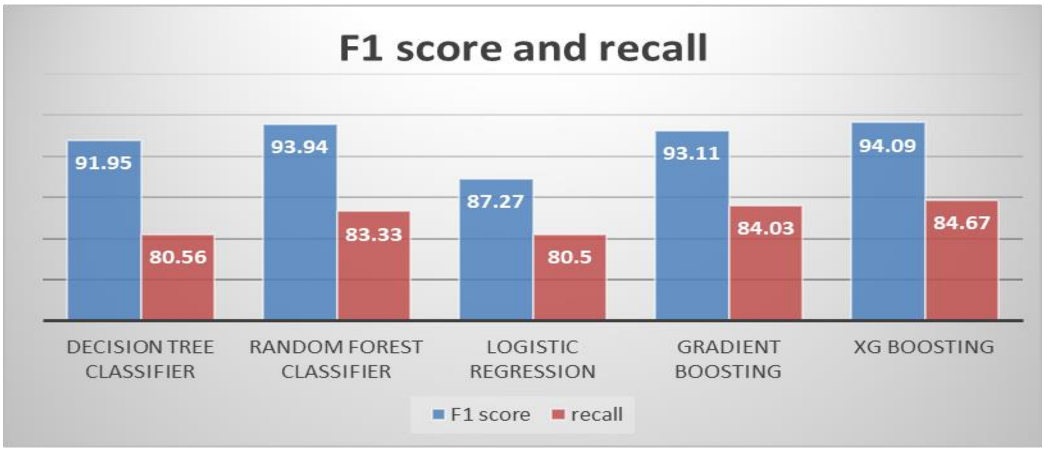
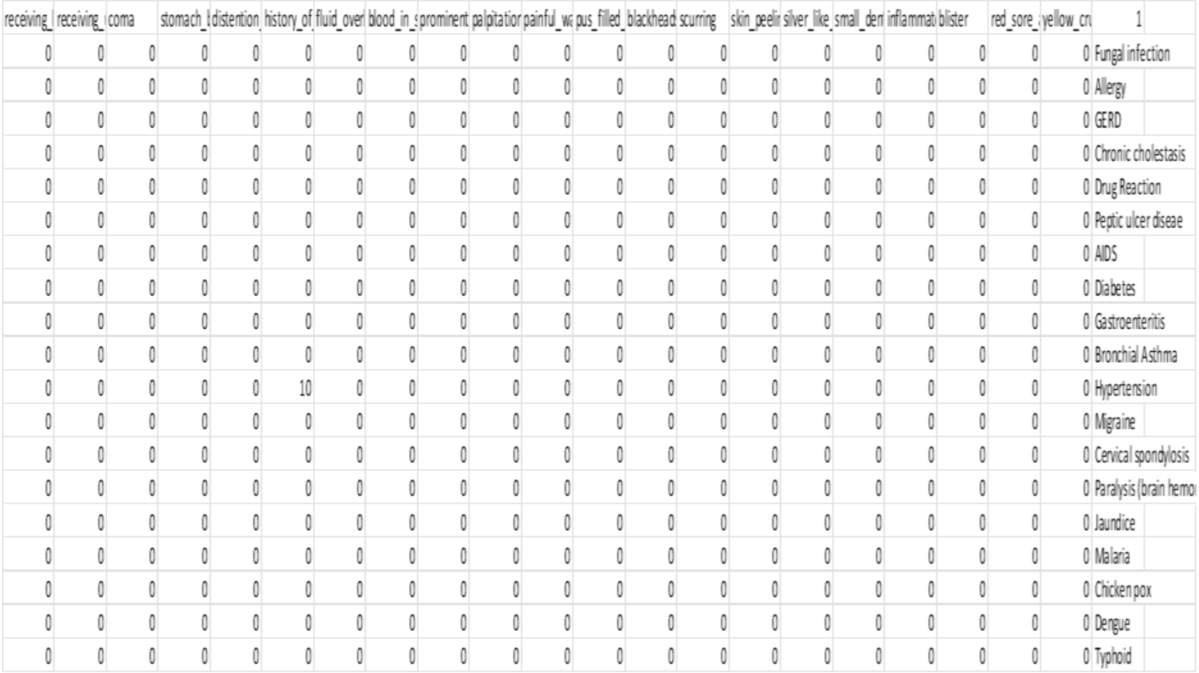


Fig.3 Barchart of F1 Score and Recall



receiving_	receiving_	coma	stomach_	distention_	history_of_	fluid_over	blood_in_s	prominent	palpitation	painful_w	pus_filled_	blackhead	scurring	skin_peeli	silver_like_	small_den	inflammat	blister	red_sore_	yellow_cr	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Chronic cholestasis
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Drug Reaction
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Peptic ulcer diseae
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	AIDS
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Diabetes
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Gastroenteritis
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Bronchial Asthma
0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Hypertension
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Migraine
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Cervical spondylosis
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Paralysis (brain hemo
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Jaundice
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Malaria
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Chicken pox
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Dengue
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Typhoid

Fig.4 Dataset

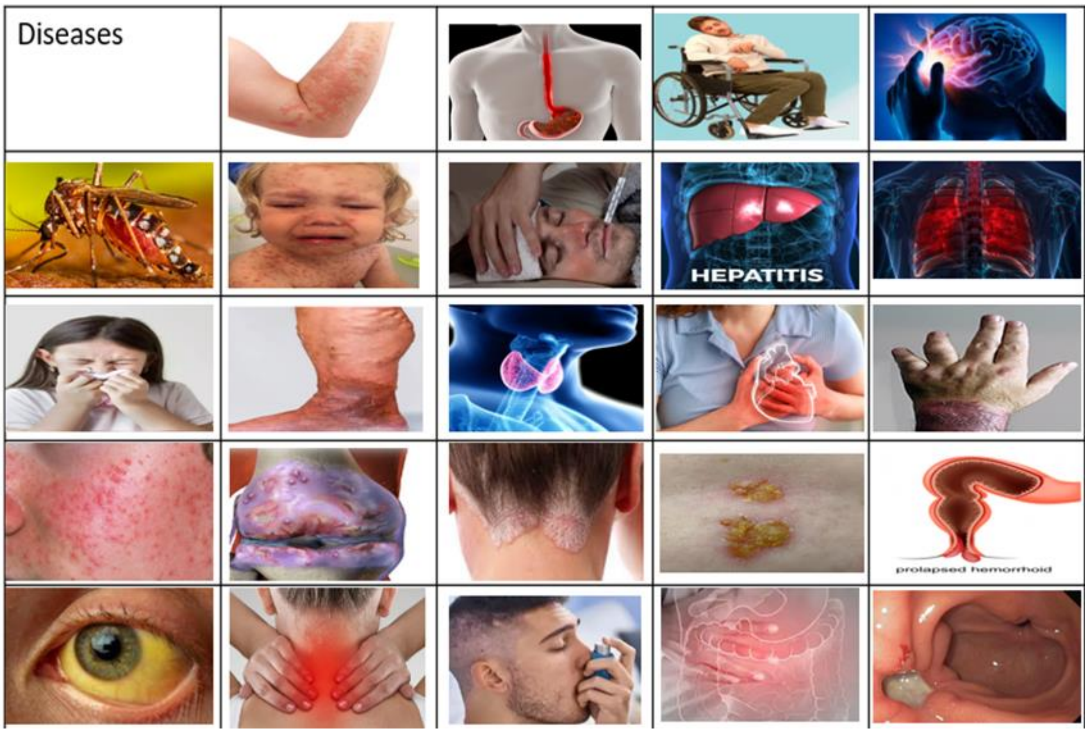


Fig.5 Diseases

Advantages of the Proposed System

1. Automation and Efficiency: Our system automates the disease prediction process using machine learning algorithms, eliminating the need for manual matching of patient records. This automation enhances efficiency and accuracy in forecasting lifestyle diseases based on user symptoms, saving time for both patients and healthcare providers.

2. Comprehensive Integration: Unlike existing systems that may lack comprehensive integration, our solution offers a centralized platform within a Flask web application. Users can access multiple medical services, including disease prediction, appointment booking, and medical information, all within one cohesive interface. This integration streamlines workflows and enhances user experience.

3. Accessibility and Convenience: By incorporating features such as an AI-powered chatbot and online appointment scheduling, our system improves accessibility and convenience for users. Patients can easily seek disease-related information, schedule appointments with doctors, and even purchase medications online, reducing the need for physical visits to healthcare facilities.

4. Proactive Health Management: Our platform's utilization of data logs to identify virus-affected areas enables proactive health management. By analyzing patterns and trends in health data, the system can alert users to potential risks in their vicinity, empowering them to take preventive measures such as maintaining social distance.

IV. CONCLUSION

In conclusion, the comparison of our proposed system with existing technologies highlights to build a Ml algorithm that accurately predict the diseases with five symptoms input and also help the customer in booking and buying medicine and help him get in contact with doctor.

In this application machine learning algorithms like decision tree, logistic regression, random forest and boosting algorithms as gradient boosting, xgboosting based on this algorithms decision tree algorithm will perform accurately with 97% based on medical helper application is developed by testing disease dataset with multiple machine learning algorithms and most accurate algorithm is used to predict disease which is used in flask web framework. Using this framework health website is designed which has doctor appointment booking, chat bot helper, medicine booking, disease prediction all health-related services are integrated in single application.

Our ML-based medical helper application leverages a carefully selected machine learning algorithm to accurately predict diseases based on six input symptoms. Integrated within the Flask web framework, this application offers a comprehensive health website featuring doctor appointment booking, chatbot assistance, medicine purchasing, and disease prediction functionalities.

By consolidating these health-related services into a single platform, we provide users with a seamless and efficient means of accessing medical assistance and support, ultimately enhancing the overall healthcare experience and promoting proactive health management.

ACKNOWLEDGEMENT

This project includes expanding the dataset to include more diverse and complex medical conditions, enhancing the model's robustness. Incorporating advanced algorithms like random forests or deep learning could further improve accuracy and predictive capabilities. Additionally, integrating real-time data from wearable devices and electronic health records can make the model more applicable in clinical settings. Finally, collaborating with healthcare professionals to validate and refine the model can ensure its practical utility in diagnosing and managing diseases.

our platform's utilization of data logs to identify virus-affected areas further enhances its proactive approach to disease management, setting it apart as a comprehensive and innovative solution in the realm of healthcare technology.

V. REFERENCES

[1] D. V. K, T. K. Ramesh and S. A, "A Machine Learning based Ensemble Approach for Predictive Analysis of Healthcare Data," 2020 2nd PhD Colloquium on Ethically Driven Innovation and Technology for Society (PhD EDITS), 2020, pp. 1-2, doi: 10.1109/PhDEDITS51180.2020.9315300.

[2] V. Kumar, D. R. Recupero, D. Riboni and R. Helaoui, "Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification From Clinical Notes," in IEEE Access, vol. 9, pp. 7107-7126, 2021, doi: 10.1109/ACCESS.2020.3043221.

[3] S. S. Ayachit, T. Kumar, S. Deshpande, N. Sharma, K. Chaurasia and M. Dixit, "Predicting H1N1 and Seasonal Flu : Vaccine Cases using Ensemble Learning approach," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020,

pp. 172-176, doi: 10.1109/ICACCCN51052.2020.9362909.

[4] Nafiz Imtiaz Khan, Tahasin Mahmud, Muhammad Nazrul Islam, and Sumaiya Nuha Mustafina. 2020. Predictionof Cesarean Childbirth using Ensemble Machine Learning Methods. In <i>Proceedings of the 22nd International Conference on Information Integration and Web-based Applications &amp; Services</i> (<i>iiWAS '20</i>). Association for Computing Machinery, New York, NY, USA, 331–339. DOI:https://doi.org/10.1145/3428757.3429138.

[5] R. MurtiRawat, S. Panchal, V. K. Singh and Y. Panchal, "Breast Cancer Detection Using K-Nearest Neighbors, Logistic Regression and Ensemble Learning," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 534-540, doi: 10.1109/ICESC48915.2020.9155783.

[6] W. Wang, J. Lee, F. Harrou and Y. Sun, "Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning," in IEEE Access, vol. 8, pp.

[7] A.Shah, D. Lalakiya, S. Desai, Shreya and V. Patel, "Early Detection of Alzheimer's Disease Using Various Machine Learning Techniques: A Comparative Study," 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), 2020, pp. 522-526, doi: 10.1109/ICOEI48184.2020.9142975.

[8] Saloni Kumari, Deepika Kumar, Mamta Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier", International Journal of Cognitive Computing in Engineering, Volume2, 2021.

[9] Ibrahim I., Abdulazeez A. The role of machine learning algorithms for diagnosing diseases. Journal of Applied Science and Technology Trends . 2021;2(1):10–19. doi: 10.38094.jastt20179.

[10] Jain D., Singh V. Feature selection and classification systems for chronic disease prediction: a review. Egyptian InformaticsJournal . 2018;19(3):179189.doi: 10.1016/j.eij.2018.03.002.

[11] Soni V. D. Chronic disease detection model using machine learning techniques. International Journal of Scientific & Technology Research . 2020;9(9):262–266.

[12] Ge R., Zhang R., Wang P. Prediction of chronic diseases with multi-label neural network. *IEEE Access* . 2020;8:138210–138216. doi: 10.1109/access.2020.3011374.