International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

xx

COPY RIGHT

ELSEVIER SSRN

IJIEMR Transactions, online available on 29th Dec 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-13&issue=Issue4

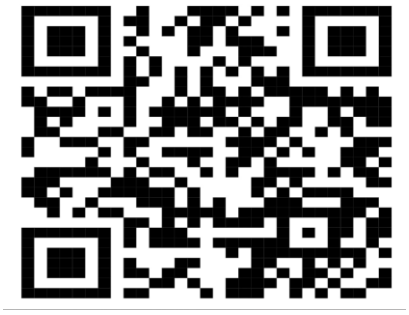
10.48047/IJIEMR/V13/ISSUE 04/62

TITLE: HEALTH MONITORING AND DISEASE DETECTION USING MACHINE LEARNING

Volume 13, ISSUE 04, Pages: 557-565

Paper Authors **G. VANI, K. KARUNAKAR, CH. VAMSHIDHAR REDDY, K. SREEKAR REDDY, A. VENKAT REDDY**

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER



To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

HEALTH MONITORING AND DISEASE DETECTION USING MACHINE LEARNING

1*G. VANI, 2*K. KARUNAKAR, 3.CH. VAMSHIDHAR REDDY, 4.K. SREEKAR REDDY, 5.A. VENKAT REDDY

1,2,3,4,5 DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, SREENIDHI INSTITUTE OF SCIENCE AND TECHNOLOGY, TELANGANA, INDIA

***Abstract*—** *Health monitoring and disease detection are critical components of modern healthcare, with machine learning serving as a cornerstone in predictive analysis. While previous studies have predominantly focused on individual diseases, there is a need for a unified system capable of predicting multiple diseases simultaneously. This paper introduces a comprehensive health monitoring and disease detection website that leverages machine learning algorithms, including Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Naive Bayes, and logistic regression, to enhance the accuracy and performance of disease prediction. By harnessing the Flask micro web framework for Python, our system facilitates the prediction of various diseases, such as diabetes, heart disease, liver disease, and kidney disorders, within a single platform. To evaluate the performance of our proposed system, we employ a diverse dataset and extract essential features for disease analysis. Our approach utilizes machine learning algorithms to efficiently predict multiple diseases, with Flask Python pickling preserving model behavior for seamless deployment and scalability. This research underscores the significance of analyzing a wide range of diseases, enabling comprehensive patient monitoring and early warning systems to reduce mortality rates. By advancing machine learning in healthcare, our study demonstrates the effectiveness of our approach in predicting multiple diseases compared to benchmark methods. The integration of various machine learning techniques and the Flask framework highlights the potential for future developments in disease prediction and monitoring.*

Keywords*—Machine learning, Flask, Decision Tree, Support Vector Machine (SVM), Random Forest, Logistic Regression, K-Nearest Neighbors (KNN).*

I. INTRODUCTION

The healthcare business faces huge data collection and processing challenges due to the massive volume of multidimensional data generated by a variety of sources, including clinical parameters, patient records, diagnostic information, and medical devices. This huge amount of data requires sophisticated processing and evaluation to yield significant insights for informed decision-making. Medical data mining has emerged as an important technique for finding hidden patterns in large datasets, and various data mining technologies and machine learning methodologies are being applied to transform healthcare systems.

By analyzing large databases, medical data mining enables the identification of significant patterns, correlations, and relationships among numerous variables. It serves as a valuable instrument in healthcare, facilitating the gathering, organization, and systematic analysis of patient data. This technology combines multiple analytic methodologies with complex algorithms to explore massive datasets, identifying inefficiencies, best practices, and opportunities for improvement in healthcare delivery.

Medical data mining is important in disease diagnosis, treatment, and prevention since it helps with early detection, provides insights for tailored medication, and improves medical practitioners' understanding of fundamental mechanisms in the medical domain. However, diagnosing chronic illnesses remains difficult, relying heavily on clinical judgment based on symptoms and physician experience. As medical systems evolve and new therapies are introduced, healthcare personnel are finding it increasingly difficult to stay up to date on diagnostic guidelines.

To address these concerns, merging machine learning techniques with medical expertise has prompted increasing interest in automating the diagnostic process. Machine learning algorithms surpass skilled clinicians in terms of diagnostic accuracy. These methods collect information from disease datasets to help in diagnosis, prediction, prevention, and treatment.

II. BASIC CONCEPTS

Decision Tree:

Decision trees are versatile machine learning models that mimic human decision-making processes. They are organized in a tree-like structure, with each internal node indicating a feature-based choice, each branch representing the outcome of that decision, and each leaf node representing the final conclusion or prediction. Decision trees are particularly useful for classification and regression tasks, and they are easy to understand, making them popular in a wide range of applications.

Support Vector Machine (SVM):

SVM is a powerful supervised learning approach for classification and regression analysis. It works by selecting the optimum hyperplane for categorizing data points while maximizing the margin between classes. SVM works well in high-dimensional environments, and it is especially beneficial when dealing with linearly separable data or data with complex decision boundaries due to the usage of kernel functions.

Random Forest:

Random Forest is an ensemble learning strategy that use many decision trees to increase prediction accuracy while reducing overfitting. During training, it generates a forest of decision trees, each trained on a randomly selected subset of the data and features. During prediction, the output from many trees is averaged or combined to get the final forecast. Random Forest is often used for classification and regression applications due to its resistance to noise and outliers.

Logistic Regression (LR):

Despite its name, logistic regression is a linear model that solves binary classification problems. It approximates the probability of a binary result by applying a logistic function to the input features. Logistic regression is simple, effective, and understandable, making it a common choice for binary classification problems when the relationship between features and outcomes is assumed to be linear.

K-Nearest Neighbors (KNN):

KNN is a simple classification algorithm that makes predictions based on the majority vote of the k nearest data points in the feature space. It does not need training because it saves all available data points and labels, making it a lazy learning algorithm. KNN is versatile and may be used for both classification and regression applications; nevertheless, it may suffer from the curse of dimensionality and be computationally expensive on large datasets.

III. LITERATURE SURVEY

Prabal Verma et al. [1] describe a cloud-centric IoT-based illness detection healthcare platform intended for mobile health applications. Using IoT advancements, the framework forecasts potential diseases and their severity levels before moving data analysis to the cloud for improved processing capability. They define key keywords for creating user-centric health metrics and provide an architectural prototype for smart student healthcare. The approach improves disease prediction accuracy over baseline methods by methodically producing data and using classification algorithms. This study focuses on the potential of cloud-based IoT systems to improve healthcare diagnostics.

Osman Salem et al. [2] describe a lightweight solution for online anomaly detection in medical wireless body area networks (WBANs) in the paper "Online Anomaly Detection in Wireless Body Area Networks for Reliable Healthcare Monitoring." The system, which employs a smartphone as a base station, focuses on detecting inaccurate measurements in real-time WBAN data while accounting for the smartphone's restricted capabilities. By combining Haar wavelet decomposition, nonseasonal Holt-Winters forecasting, and the Hampel filter, the system finds abnormalities accurately while avoiding false alarms. Results from real physiological datasets demonstrate that the system is both accurate and efficient for real-time diagnosis in healthcare monitoring applications.

Malik Bader Alazzam et al. [3] provide a machine learning implementation of a diabetic patient monitoring system in their work "Machine Learning Implementation of a Diabetic Patient Monitoring System Using Interactive E-App." The goal is to improve self-management and avoid type 2 diabetes by implementing lifestyle changes. The study details the creation, implementation, and testing of a diabetes self-management smartphone app that monitors nutritional intake and health data. The software collects Bluetooth movement data from wearable insole devices to track carbohydrate intake, blood glucose levels, medication adherence, and physical activity. Two machine learning models, SVM and decision tree, distinguish sitting and standing positions with 86% accuracy. The decision tree model is then used in a real-time activity classification system. This study emphasizes the increasing importance of mobile health self-management apps in the treatment of chronic disorders.

Vaneeta Bhardwaj et al. [4] discuss an IoT-based smart health monitoring system for COVID-19 in their paper "IoT-Based Smart Health Monitoring System for COVID-19." The device uses IoT technology to track vital signs such as blood pressure, heart rate, oxygen level, and temperature, reducing the need for frequent doctor visits. It improves communication between local clinics and city hospitals by sending out alerts for deviations from standard values. The system, which has comparable accuracy to commercial systems, allows for real-time data collection and storage for informed decision-making, resulting in early COVID-19 identification and treatment.

Zulfiqar Ali et al. [5] offer an automated health monitoring system for patients experiencing voice issues in smart cities. The system aims to enhance healthcare in smart cities by utilizing advancements in IoT and cloud computing, especially for the growing elderly population. Early detection is necessary for voice problems, which have an impact on both personal and professional lives. The technology detects abnormalities in speech output using linear prediction analysis, and it achieves high accuracy rates of 99.94% for speech with sustained vowels and 99.75% for speech that flows. The method is perfect for real-world use since it can effectively discern between normal and disordered voices by focusing on lower frequencies.

Moeen Hassanalieragh et al. [6] analyze the possibilities of health monitoring and management using Internet-of-Things (IoT) sensing and cloud-based processing in their paper "Health Monitoring and Management Using Internet-of-Things (IoT) Sensing with Cloud-Based Processing: Opportunities and Challenges". They underline the transformative potential of IoT-enabled healthcare, such as disease prevention, personalized treatment alternatives, and cost reductions. However, the study acknowledges the difficulties in accomplishing this goal, underlining the need for further research and improvement in this area.

Shadman Nashif et al. [7] propose a heart disease detection system employing machine learning algorithms and a real-time cardiovascular health monitoring system." The study describes a cloud-based cardiac disease prediction system that uses the SVM algorithm with good accuracy. Furthermore, a real-time patient monitoring system is created with Arduino, capable of sensing numerous parameters and relaying data to a central server for doctor visualization. When patient parameters surpass criteria, the device sends an alarm to doctors using GSM technology, allowing for rapid medical intervention.

Choi et al. [8] present a systematic analysis of web-based infectious disease surveillance systems in their paper "Web-based infectious disease surveillance systems and public health perspectives: a systematic review." Recognizing the importance of early detection and response in infectious disease management, the study investigates the impact of emerging web-based data sources in boosting established monitoring systems. A systematic review framework was applied to literature published between 2000 and 2015 to evaluate eleven surveillance systems for their frequency of use and critical attributes. The evaluation focuses on the advantages of web-based surveillance systems over traditional systems, such as their intuitiveness, versatility, low cost, and real-time operation.

Kamble et al. [9] describe an IoT-based Patient Health Monitoring System with Nested Cloud Security in their paper "IoT-based Patient Health Monitoring System with Nested Cloud Security." Taking use of advances in the Internet of Things (IoT), the suggested system monitors patients' physiological data such as temperature, heartbeat, and ECG using sensors attached to a Raspberry Pi Board. Data transfer to a local server allows for real-time monitoring, with caregivers and medical professionals receiving immediate notification of any significant signs or symptoms via messages or emails. To ensure data security, Shamir's Secret Key Sharing Algorithm is used, in which data and keys are divided and stored in separate locations, with threshold cryptography used for retrieval. Furthermore, the system uses the support vector machine (SVM) algorithm for cardiac illness.

Nosirov et al. [10] describe a real-time multi-parametric human health monitoring and prediction system in their paper titled "Real-time multi-parametric human health monitoring and prediction system." The system uses cutting-edge technology, such as

the Internet of Things (IoT) and mobile devices, to allow patients and healthcare providers to communicate and monitor each other remotely. It analyzes clinical data to identify specific conditions for early detection or awareness, providing originality, multifunctionality, and cost in healthcare monitoring.

IV. METHODOLOGY

Data Collection:

The initial phase in our process is to collect datasets related to numerous ailments, including cancer, diabetes, heart disease, renal disease, liver disease, malaria, and pneumonia. These databases include attributes connected to patient health metrics, making it easier to analyze and forecast disease outcomes.

The breast cancer dataset comprises 32 features extracted from digitized fine needle aspirate (FNA) images of breast tumors. These features encompass characteristics such as size, texture, shape, and cell nuclei properties. The diabetes dataset consists of nine attributes related to health indicators, including factors like glucose levels, blood pressure, skin thickness, and body mass index (BMI), alongside demographic information such as age and pregnancy history. The Fetal Health dataset encompasses 21 variables about fetal health monitoring during pregnancy, capturing parameters like fetal heart rate, uterine contractions, and accelerations. For the heart dataset, 14 attributes are recorded, including patient demographics, medical history, and diagnostic test results such as electrocardiographic measurements and exercise tolerance. The Indian liver patient dataset contains 11 features associated with liver disease diagnosis, encompassing biochemical markers like bilirubin levels and liver enzyme tests. In the kidney disease dataset, 25 attributes are included, covering factors like patient demographics, clinical measurements, and medical history indicators such as hypertension and diabetes status. Lastly, the stroke dataset comprises 12 attributes, including demographic information, lifestyle factors, and clinical measurements, with a target variable indicating the occurrence of a stroke event.

Preprocessing and Model training:

The collected data undergoes preprocessing to ensure its quality and suitability for machine learning algorithms. This includes data cleaning to remove null values and redundant attributes, data imputation to handle missing values, and normalization to scale the data appropriately.

Preprocessing the breast cancer dataset entails deleting extraneous columns, such as the "Unnamed" column, and converting the categorical variable "diagnosis" to numerical values with LabelEncoder. The dataset is then separated into input features (X) and target variables (y). A pair plot from the Seaborn library is used to visually explore data, displaying correlations between various variables and cancer diagnosis. The dataset is then divided into training and testing sets using the train_test_split function. The dataset was trained using the Random Forest classifier, which uses ensemble learning to create several decision trees and aggregate their predictions. The model's performance was then evaluated using the accuracy metric.

The Indian Liver Patient dataset is loaded and analyzed, showing details such as missing values in the "Albumin_and_Globulin_Ratio" column. Data visualization tools, such as count plots and pair plots, are used to analyze the prevalence of liver disease by gender and age group. The next step is to use scatter plots and correlation matrices to detect relationships between various parameters such as bilirubin levels, liver enzymes, and proteins. Features with high correlations are detected, implying the possibility of removing duplicate variables. Categorical data, such as gender, is converted to numerical form for model compatibility. The data is then split into training and testing sets, features are scaled, and dimensionality reduction is applied using Principal Component Analysis (PCA). Several classification techniques, such as Logistic Regression, Gaussian Naive Bayes, Random Forest, and K-Nearest Neighbors, are trained and tested with accuracy scores, confusion matrices, and classification reports.

Preprocessing the stroke dataset entails removing the 'id' column and handling missing values, investigating the distribution of the target variable ('stroke'), and generating descriptive statistics for the dataset. Further analysis entails inspecting the data types, examining the distribution of categorical variables such as 'work_type' using count plots, and converting categorical variables into numerical codes for model training. Furthermore, dictionaries are used to convert categorical variables into numerical representations. Moving on to model development, the dataset is divided into features (X) and target variables (Y), and

preprocessing activities such as feature scaling with MinMaxScaler and train-test splitting are carried out. Then, a RandomForestClassifier is used to train the model, and its performance is measured using an accuracy score.

The kidney disease dataset is preprocessed to remove missing values, fix incorrect classifications, and convert categorical variables to numerical representations. Additional preprocessing stages include removing extraneous columns, switching data types, and encoding categorical variables with predefined dictionaries. Exploratory data analysis (EDA) employs visualizations such as heatmaps to better understand the relationship between features and the goal variable (classification). The dataset is divided into features (X) and target variables (y), and training and testing sets are created with train_test_split. A Random Forest Classifier model is trained on the training data and assessed using accuracy measures like the confusion matrix and accuracy score.

After loading the cardiac dataset, exploratory data analysis (EDA) is used to better comprehend its structure, display distributions, and investigate feature relationships. The dataset is then preprocessed, which includes managing missing values, scaling the features with StandardScaler, and dividing them into training and testing sets. Several machine learning models, such as Random Forest, K-Neighbors Classifier, and Support Vector Classifier (SVC), are trained and tested for heart disease prediction. Each model's performance is assessed using evaluation measures such as the confusion matrix, accuracy score, and classification report.

The diabetes dataset is loaded, preprocessed to remove any missing values, and divided into training and testing datasets. Then SVM and Random Forest classifiers are created. The model is then trained on the training data using the fit approach to discover the underlying patterns and correlations between the input variables and the diabetes outcomes.

V. RESULTS

The Random Forest classifier was used to train each dataset, and its performance was measured using accuracy measures. Across all datasets, Random Forest demonstrated strong predictive powers. In the cancer dataset, Random Forest attained an accuracy of 94%, demonstrating its ability to distinguish between malignant and benign tumors. In the heart dataset, the Random Forest classifier achieved 85% accuracy, indicating its ability to diagnose heart-related illnesses. Furthermore, in the renal dataset, Random Forest achieved an extraordinary accuracy of 99%, demonstrating its exceptional capacity to reliably identify kidney problems. However, with the liver dataset, Random Forest achieved a significantly lower accuracy of 78%, implying that more refining may be required for liver disease prediction. Finally, in the diabetes dataset, Random Forest fared remarkably well, with an accuracy of 92%, demonstrating its efficacy in predicting diabetic health outcomes. Overall, the Random Forest classifier's constant high accuracies across varied datasets demonstrate its reliability and applicability for illness prediction tasks.

Flask was used to create a graphical user interface (GUI) that allows users to interact and streamlines the process of collecting disease predictions. The GUI offers consumers an easy-to-use platform for entering relevant medical factors and receiving forecasts for a variety of conditions. The GUI improves accessibility and empowers people to make informed health decisions by combining intuitive design with seamless functionality. Users can quickly navigate through the GUI's various sections, including those for diabetes, cancer, heart disease, liver disease, renal problems, stroke, malaria, and pneumonia prediction. Figures 1, 2, 3, 4, 5, and 6 depict unique disease classifications matching to various input conditions in the graphical user interface (GUI). Each graphic depicts the outcome of the disease prediction process based on the user's specified input parameters. The GUI improves the entire user experience by combining user-friendly features and interactive aspects, allowing for more effective illness prediction and monitoring.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

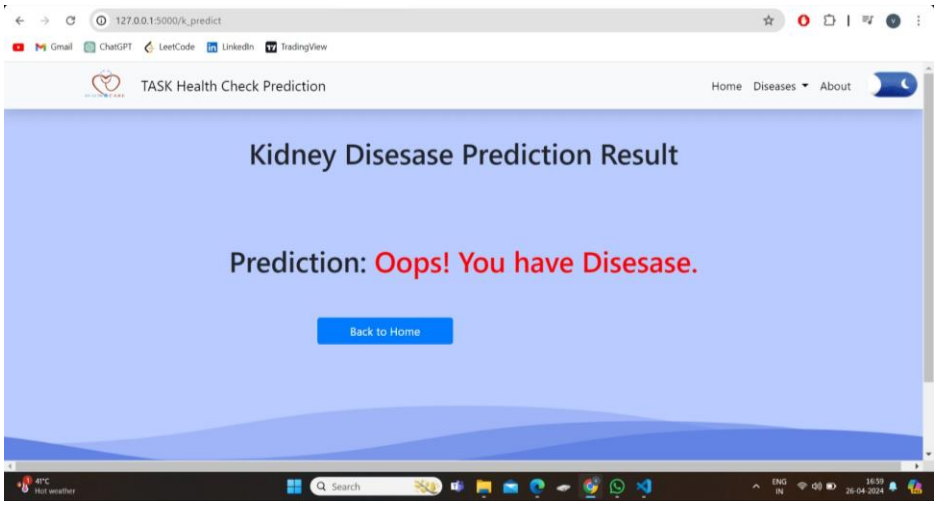


Figure 1: Visualization of kidney prediction outcome based on input parameters

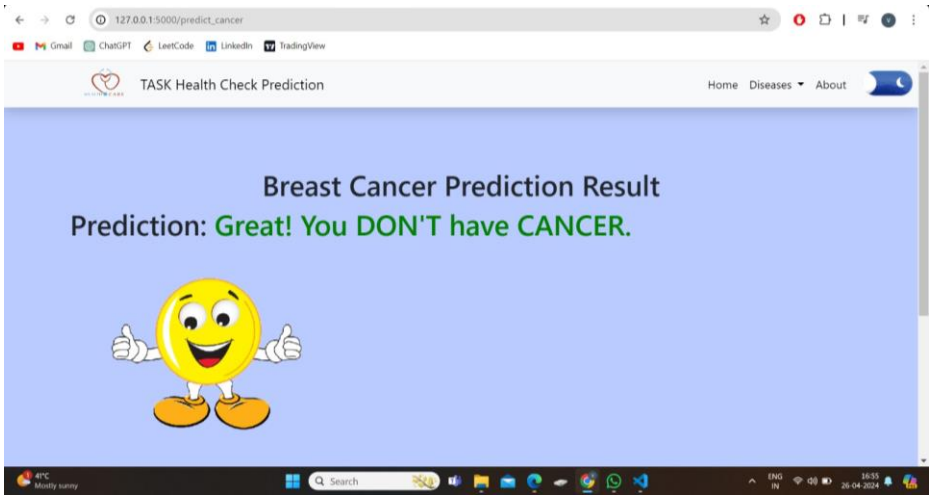


Figure 2: Visualization of cancer prediction outcome based on input parameters

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

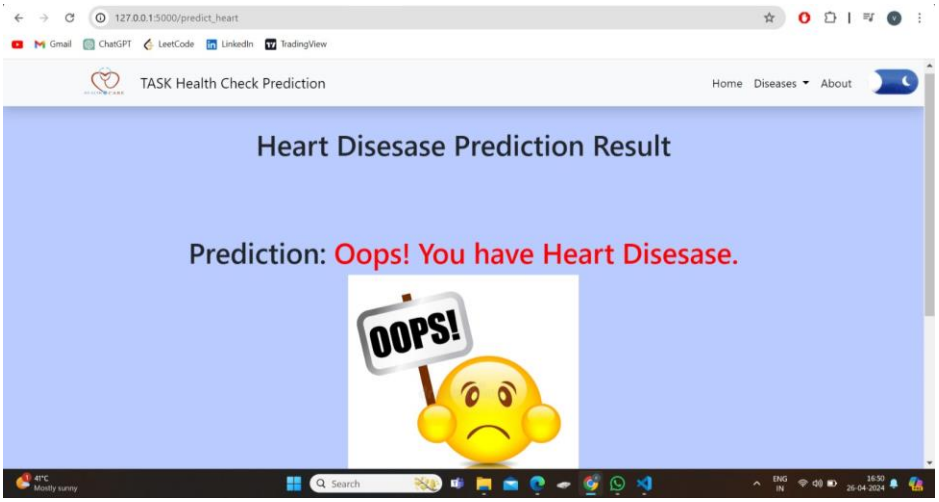


Figure 3: Visualization of heart disease prediction outcome based on input parameters

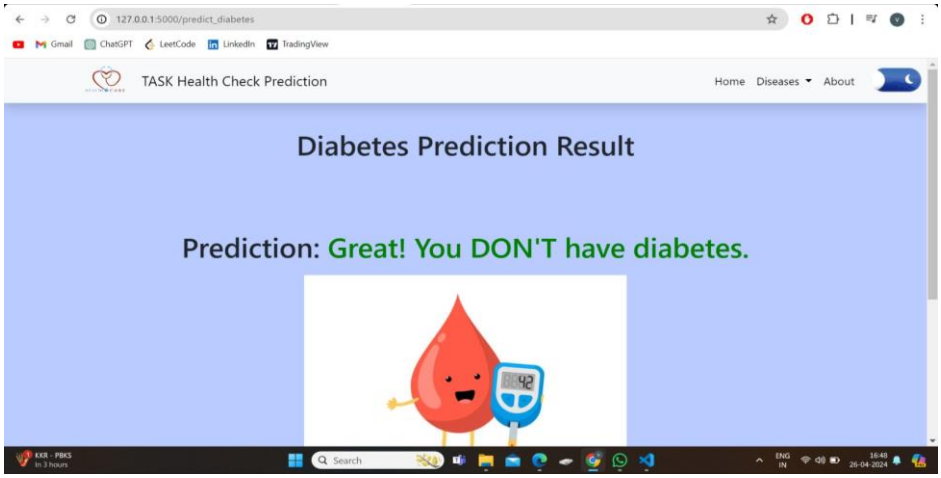


Figure 4: Visualization of diabetes prediction outcome based on input parameters

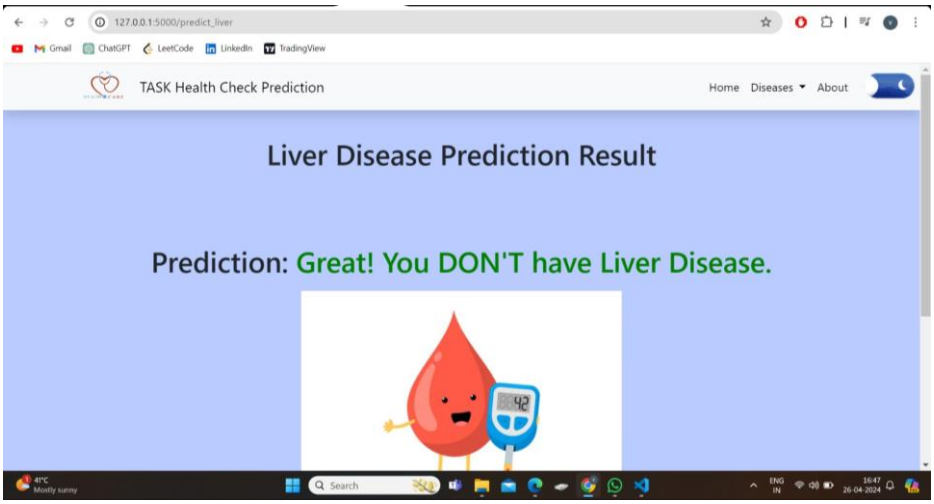


Figure 5: Visualization of liver disease prediction outcome based on input parameters

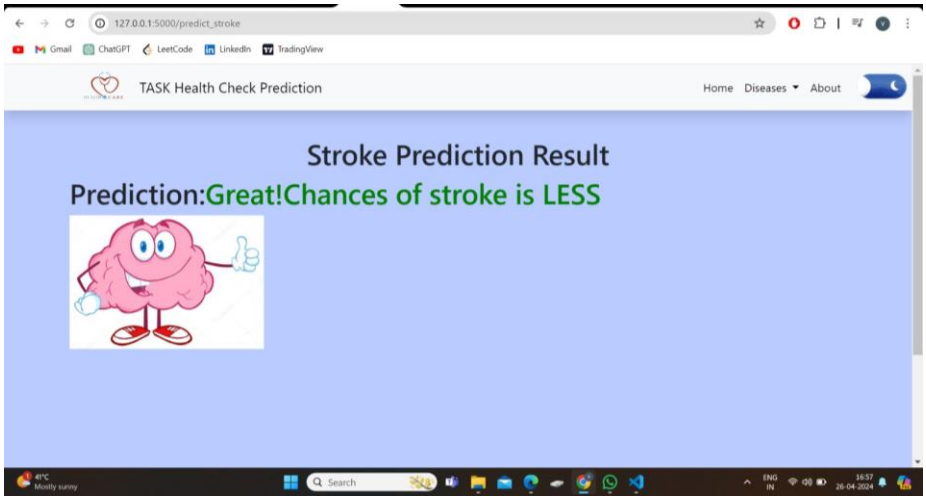


Figure 6: Visualization of stroke prediction outcome based on input parameters

VI. CONCLUSION

Our findings not only show the usefulness of machine learning algorithms in disease prediction, but also emphasize the importance of multidisciplinary collaboration between the healthcare and data science fields. By bridging the gap between these disciplines, we may use data-driven insights to improve patient care and outcomes. We used different machine learning methods, including Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Naive Bayes, and logistic regression, to show that our approach was effective in predicting multiple ailments, including diabetes-related heart disease. To increase disease prediction accuracy, we extracted essential traits from many datasets through rigorous feature engineering and exploratory data analysis. Our comprehensive health monitoring and disease detection website, built on the Flask micro web framework, provides a unified platform for forecasting conditions such as diabetes, heart disease, liver disease, and kidney diseases.

VII. FUTURE WORK

Our findings identified several areas for future research and development in healthcare and disease prediction. Machine learning algorithms can increase the accuracy of sickness diagnosis. Deep learning, ensemble learning, and transfer learning approaches can improve accuracy by discovering complex patterns in healthcare data. Our project's expansion to include real-time disease monitoring and continuous data collection offers the opportunity to create a more dynamic prediction model. By merging wearable devices, Internet of Things (IoT) sensors, and health records, the system may continuously analyze data and provide timely alerts for early disease detection. Integrating the disease prediction model with electronic health records (EHR) systems allows for a fuller view of a patient's medical history. Using a broader set of patient data, such as medical diagnoses, test results, and medication history, can increase disease forecasting accuracy and dependability. Overall, these prospective pathways demonstrate the potential for machine learning to improve disease prediction, early intervention, and tailored healthcare, leading to better patient outcomes and community health management.

REFERENCES

[1] P. Verma and S. K. Sood, “Cloud-centric IoT based disease diagnosis healthcare framework,” Journal of Parallel and Distributed Computing, vol. 116, pp. 27–38, Jun. 2018, doi: https://doi.org/10.1016/j.jpdc.2017.11.018.

[2] O. Salem, Y. Liu, A. Mehaoua, and R. Boutaba, "Online Anomaly Detection in Wireless Body Area Networks for Reliable Healthcare Monitoring," IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 5, pp. 1541–1551, Sep. 2014, doi: https://doi.org/10.1109/jbhi.2014.2312214.

[3] M. B. Alazzam, H. Mansour, F. Alassery, and A. Almulihi, "Machine Learning Implementation of a Diabetic Patient Monitoring System Using Interactive E-App," Computational Intelligence and Neuroscience, vol. 2021, pp. 1–7, Dec. 2021, doi: https://doi.org/10.1155/2021/5759184.

[4] V. Bhardwaj, R. Joshi, and A. M. Gaur, "IoT-Based Smart Health Monitoring System for COVID-19," SN Computer Science, vol. 3, no. 2, Jan. 2022, doi: https://doi.org/10.1007/s42979-022-01015-1.

[5] Z. Ali, G. Muhammad, and M. F. Alhamid, "An Automatic Health Monitoring System for Patients Suffering From Voice Complications in Smart Cities," IEEE Access, vol. 5, pp. 3900–3908, 2017, doi: https://doi.org/10.1109/ACCESS.2017.2680467.

[6] M. Hassanalieragh et al., "Health Monitoring and Management Using Internet-of-Things (IoT) Sensing with Cloud-Based Processing: Opportunities and Challenges," 2015 IEEE International Conference on Services Computing, Jun. 2015, doi: https://doi.org/10.1109/scc.2015.47.

[7] S. Nashif, Md. R. Raihan, Md. R. Islam, and M. H. Imam, "Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System," World Journal of Engineering and Technology, vol. 06, no. 04, pp. 854–873, 2018, doi: https://doi.org/10.4236/wjet.2018.64057.

[8] J. Choi, Y. Cho, E. Shim, and H. Woo, "Web-based infectious disease surveillance systems and public health perspectives: a systematic review," BMC Public Health, vol. 16, Dec. 2016, doi: https://doi.org/10.1186/s12889-016-3893-0.

[9] A. Kamble and Sonali Bhutad, "IOT based Patient Health Monitoring System with Nested Cloud Security," Dec. 2018, doi: https://doi.org/10.1109/ccaa.2018.8777691.

[10] Kh. Nosirov et al., "Real-time multi parametric human health monitoring and prediction system," Developments of Artificial Intelligence Technologies in Computation and Robotics, Aug. 2020, doi: https://doi.org/10.1142/9789811223334_0077.