Streamlit-Powered Comprehensive Health Analysis and Disease Prediction System

[1st] Mahendran.K
Department of Electronics and Communication Engineering,
Saveetha Engineering college
Chennai,TamilNadu,India
mahendrank@saveetha.ac.in

[2nd] Surya.S
Department of Electronics and Communication Engineering,
Saveetha Engineering college
Chennai,TamilNadu,India
*suryasagadevan17@gmail.com

[3rd] Thejashrayal.E
Department of Electronics and Communication Engineering,
Saveetha Engineering college
Chennai,TamilNadu,India
thejashrayal208@gmail.com

2023 International Conference on Emerging Research in Computational Science (ICERCS) | 979-8-3503-5976-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICERCS57948.2023.10434221

***Abstract*—Maintaining well-being is vital, and healthcare services play a pivotal role in society. Unlike existing AI models focusing on individual disease prediction, our goal is to create a unified platform using the Streamlit Python module, predicting multiple diseases. Employing machine learning algorithms like random forest, Logistic Regression, SVM classifiers, and Extra Tree Classifier, we identify the algorithm with the highest accuracy during the dataset training phase. This approach allows for the examination of conditions such as diabetes, heart disease, and Parkinson's disease, considering parameters like pulse rate, cholesterol levels, blood pressure, and heart rate. The project's scope extends to encompassing additional chronic illnesses and skin conditions. By leveraging core health indicators, this study demonstrates the potential to predict a broad spectrum of diseases. The significance lies in early detection and intervention, contributing to a reduction in mortality rates.**

Keywords—Healthcare Prediction, Multiple Disease Prediction, Streamlit Platform, Machine Learning Algorithms.

I. INTRODUCTION

This paper leverages machine learning to develop predictive models for diabetes, heart disease, and Parkinson's disease, addressing global health challenges stemming from sedentary lifestyles and viral hepatitis, which cause 1.34 million deaths annually. Using the Streamlit Python framework, our method creates a user-friendly interface to manage multidimensional patient data, enhancing healthcare management. By employing medical data mining with machine learning tools, the approach uncovers hidden patterns, aiding decision-making, and plays a vital role in early disease detection, prevention, and improved diagnosis.

Unlike traditional health analysis systems focusing on individual conditions, our design concurrently assesses various conditions, including diabetes, heart disease, and Parkinson's disease. This comprehensive approach allows for flexibility and future expansion to cover additional conditions. The model considers various factors, ensuring a thorough analysis and detection of diseases, even when symptoms overlap. Developers can easily integrate new diseases by incorporating associated model files, streamlining disease coverage expansion.Addressing current deficiencies in single-disease approaches, our project emphasizes adaptability, enabling a comprehensive examination of various health conditions. The model takes into account diverse parameters, facilitating nuanced disease detection. Utilizing employee data from Kaggle and the SVM machine learning algorithm with an 80-20 data split for training and validation, our strategic approach ensures accurate model selection and performance comparisons.

In summary, this multifaceted project introduces a more inclusive and flexible methodology for disease prediction, aligning with the broader goal of optimizing healthcare practices through advanced technologies. The intersection of machine learning, comprehensive disease analysis, and streamlined model development showcases potential transformative advancements in healthcare, contributing to early detection, effective treatment, and improved patient outcomes.

This approach highlights the importance of adaptability, as it encompasses the simultaneous assessment of multiple diseases, ensuring a comprehensive examination of various health conditions. Importantly, the model takes into account diverse parameters, enabling a nuanced and thorough detection of diseases, particularly in cases where symptoms may overlap. Developers can seamlessly integrate new diseases into the existing model by incorporating associated model files, streamlining the expansion of disease coverage. The inclusion of employee data from Kaggle in the project, coupled with the utilization of the SVM machine learning algorithm, adheres to an 80-20 data split for training and validation. This strategic approach not only facilitates the selection of the most accurate model but also promotes precise comparisons of their respective performances.In essence, the multifaceted project not only pioneers a more inclusive and flexible methodology for disease prediction but also aligns with the broader objective of optimizing healthcare practices through advanced technologies. The intersection of machine learning, comprehensive disease analysis, and streamlined model development via platforms like Streamlit showcases the potential for transformative advancements in the healthcare landscape, contributing to early detection, effective treatment, and improved overall patient outcomes.

The organization of this paper is as follows: Section 2 analyzes the various ex-isting works that use machine learning and deep learning techniques for multiple disease detection. Section 3 outlines the proposed methodology. Section 4 summarizes the key findings and makes recommendations for future research.

II. Literature Survey

This paper briefly reviews and consolidates findings from various research studies on disease prediction in healthcare. It focuses on methodologies and results obtained from notable papers in the field.

Recognizing the significance of early detection in mitigating risks, Gopisetti, Laxmi Deepthi, et al. [1] present a pioneering approach. The research they conducted introduces an inclusive Multiple Disease Prediction System, amalgamating forecasts for diabetes, heart disease, chronic kidney disease, and cancer into a cohesive user interface. The research employs various classification algorithms, including K-Nearest Neighbor, Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, and Gaussian naive bayes, to enhance disease prediction accuracy.

Akkem Yaganteeswarudu and associates et al,[2] carried out a relative study to charge the efficacity of the resolution Tree, Random Forest, and Logistic Retrogression algorithms in prognosticating multitudinousdiseases.The study reported that logistic retrogression achieved 92% delicacy for heart complaint bracket, Random Forest yielded 95% delicacy, and SVM attained 96% delicacy in cancer discovery.

Chetan Sagarnal et al, as documented in reference [3], utilized algorithms to analyze symptoms and predict diseases, achieving an accuracy rate of 95.12%. In a study conducted by NuzhatF. Shaikh [4] different visualization ways were assumed to comprehend modules, and algorithms were assimilated grounded on both delicacy and time taken. specially, the J48 algorithm surfaced as the top pantomime, scoring the loftiest delicacy at an emotional rate of 98.12%.

Rashmi G Saboji et al. [5] focused on predicting heart disease using the Random Forest Algorithm. The study compared results with Naïve-Bayes classifiers and found that Random Forest provided more accurate results, with an accuracy of 98%.

In a study by Pahulpreet Singh Kohli and platoon [6]complaint vaticination was supported through colorful engine literacy styles, similar as Logistic Retrogression, resolution Tree, Brace Vector Machine, Random Forest, and Adaptive Boosting. The exploration covered prognostications for bone cancer, Diabetes, and Heart complaint, pressing Logistic Retrogression as especially operative with rigor reaching95.71 for bone cancer,84.42 for Diabetes, and87.12 for Heart complaint.

Lambodar Jena et al. [7] concentrated on predicting chronic diseases' risks using distributed machine learning classifiers. They applied techniques such as Naïve Bayes and Multilayer Perceptron to predict Chronic Kidney Disease, achieving accuracies of 95% and 99.7%, respectively. Naganna Chetty and team [8] devised a system utilizing a fuzzy approach for disease prediction, focusing on diabetes and liver diseases. They employed methods like the KNN classifier, fuzzy c-means clustering, and fuzzy KNN classifier, achieving accuracies of 97.02% for diabetes and 96.13% for liver diseases.

Sayali Ambekar et al. [9] proposed Disease Risk Prediction using a convolutional neural network (CNN). They employed machine learning techniques, including the CNN-UDRP algorithm, Naïve Bayes, and KNN algorithm, achieving an accuracy of 82%, with Naïve Bayes playing a primary role in the outcome.

MinChen et al. [10] presented a disease prediction system utilizing diverse machine learning algorithms such as the CNN-UDRP algorithm, CNN-MDRP algorithm, Naïve Bayes, K-Nearest Neighbor, and Decision Tree. The system achieved a high accuracy of 94.8% in disease prediction.

Kothapeta, Haindavi, et al. [11] conducted research on creating a multi-disease prediction model through machine learning techniques and the Streamlit interface. Users can input a complaint, and the model builds and predicts the corresponding disease. Interestingly, even without a user-entered complaint, the model is capable of making predictions.

The research conducted by Vasavi, D., et al. [12] delves into the domain of healthcare data, which has evolved as an interdisciplinary field stemming from database statistics. The study emphasizes its significance in evaluating the effectiveness of medical interventions through the integration of data visualization and machine learning. Particularly, the focus lies on predicting heart disease in individuals with diabetes, a subgroup susceptible to diabetes-related heart disease due to the chronic nature of the condition.

The study conducted by Keerthi, M. S. P., et al. [13] presents a novel approach in the form of a Streamlit interface for the simultaneous diagnosis of three different diseases within a single platform. This research includes the possibility of employing different algorithms for logistic regression and support vector machines, aiming to achieve improved classification accuracies. Moreover, the pneumonia model could derive advantages from supplementary data, potentially resulting in the development of an independent model with improved accuracy.

In their research, Absar, Nurul, et al. [14] tackle the significant issue of heart disease, a major contributor to mortality. They employ four machine learning models (RF, DT, AB, and KNN) to predict coronary heart disease, utilizing CHSLB and Cleveland datasets. Through effective data preprocessing techniques, they enhance detection accuracy, with KNN showing exceptional performance at 100% accuracy for the CHSLB dataset and 97.82% accuracy for the Cleveland dataset. The RF, AB, and DT models also exhibit noteworthy accuracy levels for the CHSLB dataset.

The study led by Kalshetty, Jagadevi N., et al. [15] delves into the essential aspect of heart health, recognizing the pivotal role of the heart in oxygen transport throughout the body. This observation accentuates the imperative for proactive measures and predictive tools to address the growing risk of heart attacks in the younger population, making a valuable contribution to the discourse on heart health.

In this section, we evaluated the performance of machine learning models such as SVM, LR, Random Forest, and Extra Tree Classifier for predicting multiple diseases. The study aims to identify the best optimal model for disease detection by utilizing diverse datasets, including the Diabetes Dataset (PIDD), data from heart disease patients, and aggregated data on Parkinson's disease

III. METHODOLOGY

The proposed methodology for multiple disease prediction comprises four fundamental steps, each playing a crucial role in the system's effectiveness:

i. Data collection
ii. Data Pre-processing
iii. Model Building
iv. Performance evaluation

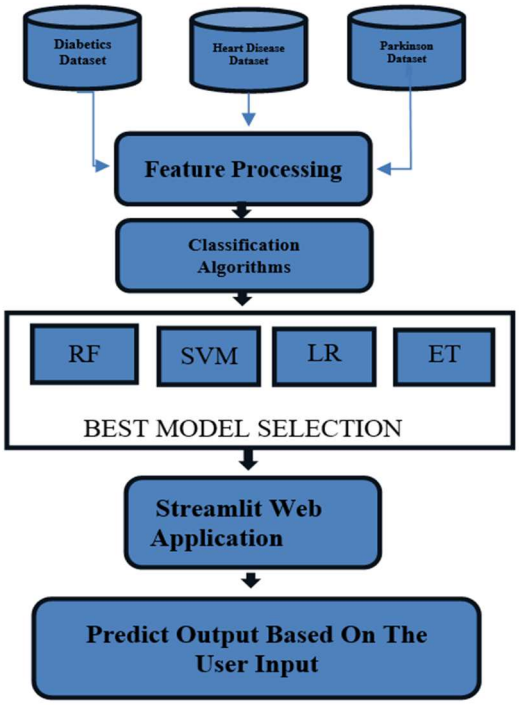

Figure 1. Flow of working methodology

Data set Collection

The data collection process for multiple disease prediction involves gathering comprehensive datasets relevant to each target disease. For instance, in the con-text of a Multiple Disease Prediction System, datasets related to diabetes, heart disease and Parkinson would be acquired. These datasets should be diverse, en-compassing instances of both healthy individuals and those affected by the re-spective diseases.

Data Preprocessing

Data pre-processing is a critical phase that involves tasks like data cleaning and noise removal. Pre-processing techniques such as data cleaning and data reduction were employed to enhance the quality and relevance of the data. The data cleaning process encompassed tasks such as addressing missing values, re-solving inconsistencies, and ensuring data integrity. Data reduction aimed to simplify the analysis by selecting a subset of the most relevant symptoms (inde-pendent variables) out of the 132 available, focusing on those closely related to the diseases of interest.

Model Building

Random Forest Classifier

Random Forest utilizes ensemble learning, a technique involving multiple algorithms or iterations of the same algorithm. In the case of Random Forest, it consists of a collection of Decision Trees. The model's generalization improves with a higher number of decision trees in the ensemble, addressing a key limitation of Decision Trees, namely overfitting.

The operational steps of Random Forest are as follows:

1.It randomly selects k symptoms from the dataset (medical records) out of m total symptoms, where k << m, to construct a decision tree.
2.This process iterates n times, resulting in n decision trees built from distinct random combinations of k symptoms or random samples of the data (bootstrap samples).
3.Each of the n decision trees takes random variables as inputs to predict diseases, storing the predictions and resulting in n predicted diseases.
4.The algorithm calculates votes for each predicted disease, selecting the mode, which represents the most frequently predicted disease, as the final prediction.

Logistic Regression

Logistic Regression is a commonly employed statistical method for binary classification tasks, aiming to predict the probability of an instance belonging to a specific class. The model employs the sigmoid (logistic) function to compress the output into a range between 0 and 1, indicative of probabilities. The sigmoid function is defined as follows:
(sigma(z) = frac{1}{1 + e^{-z}}),

In the logistic regression model, z represents a linear combination of the input features. The training process involves maximum likelihood estimation, adjusting parameters to maximize the likelihood of observed outcomes given the input features. Logistic Regression is a versatile algorithm applicable to diverse scenarios, such as medical diagnosis, spam detection, and credit scoring. Its capability to predict binary outcomes based on input features makes it suitable for a wide range of applications.

Support Vector Meachine

SVM, or Support Vector Machine, seeks to identify a hyperplane that maximally separates classes within a feature space. The optimal hyperplane is the one with the maximum margin, representing the distance between the hyperplane and the nearest data point of either class. Support vectors, the data points closest to the hyperplane, play a critical role in determining its position and orientation. Once the optimal hyperplane is established, SVM classifies new data points based on their position relative to the hyperplane. SVM is effective in high-dimensional spaces and proves particularly useful when dealing with non-linearly separable data. Its strength lies in scenarios with clear margins between classes, making it a valuable tool for various classification tasks.

Extra Tree Classifier

The Extra Trees Classifier, or Extremely Randomized Trees Classifier, is an ensemble learning algorithm used for both classification and regression tasks. In contrast to Random Forest, Extra Trees introduces an extra layer of randomness during the tree construction process..

At each split point, a random subset of features is considered for the decision, reducing sensitivity to noise in the data and enhancing diversity among individual trees. Like other decision tree-based methods, nodes in each tree are split based on specific features, but Extra Trees incorporates a randomization strategy.
The final prediction for a sample is made by aggregating predictions from all trees in the ensemble. In classification, the class with the majority of votes becomes the final prediction. The key parameter to control is the number of trees in the ensemble. This randomization strategy helps prevent overfitting, making Extra Trees potentially more robust to noise in the training data. The algorithm can also be computationally efficient due to reduced need for extensive tuning.

Performance evaluation

The accuracy of various algorithms were compared and their performance were analyzed among them the model with highest accuracy was deployed.

Models	Accuracy on diabetes	Accuracy on heart disease	Accuracy on parkinson's disease
RFC	94	78	84
SVC	82	88	87
LR	78	81	89
ETC	77	76	88

Table 1. Comparison of Accuracy of various ML models

In this table, we observe the accuracy percentages of three distinct machine learning algorithms Random Forest Classifier, Support Vector Classifier, and Logistic Regression across three specific health conditions: Diabetes, Heart Dis-ease, and Parkinson's Disease. These accuracy values provide insights into the predictive capabilities of each algorithm for their respective medical contexts. From this table. we can infere that the RFC has highest accuracy for diabetes pre-dection, SVC for heart disease prediction and LR for parkinson's disease predic-tion.

IV. Results and Discussion

The implementation of this project marks a significant leap forward in healthcare and disease prediction, offering users a host of valuable benefits. One of the most notable advantages is that users no longer need to navigate various websites or resources to obtain information about potential diseases or undergo separate assessments. This integrated system serves as a one-stop solution, sav-ing precious time and effort while providing comprehensive insights into multi-ple diseases.

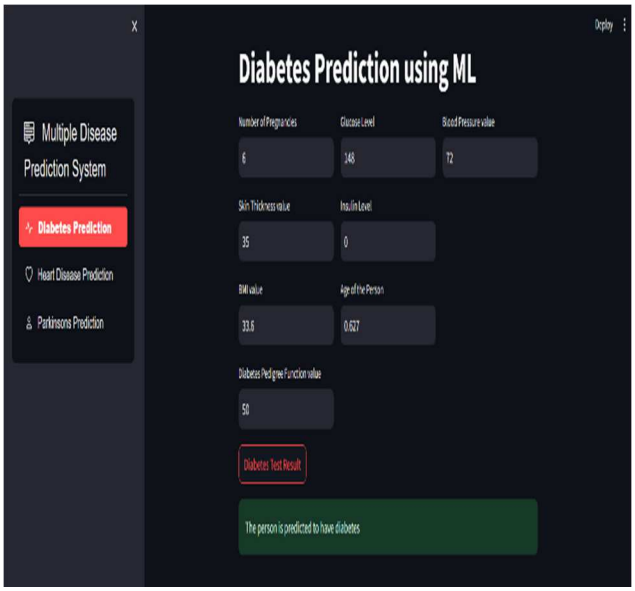

Figure.2. User interface for diabetes prediction

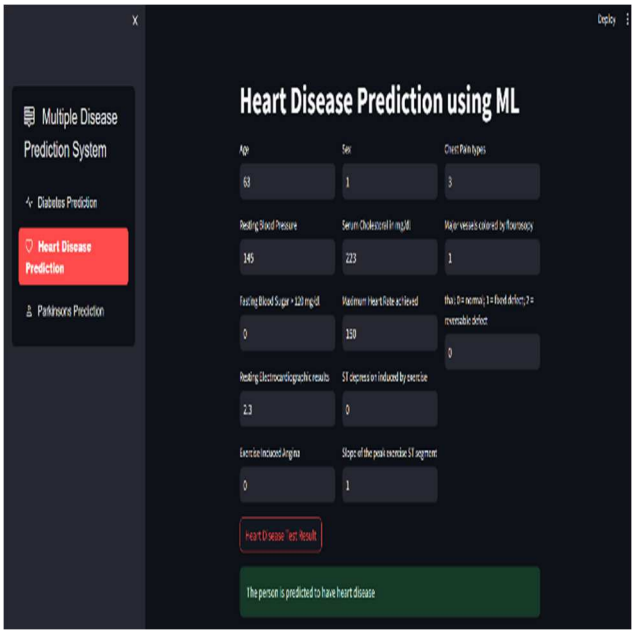

Figure.3. User interface for heart disease prediction

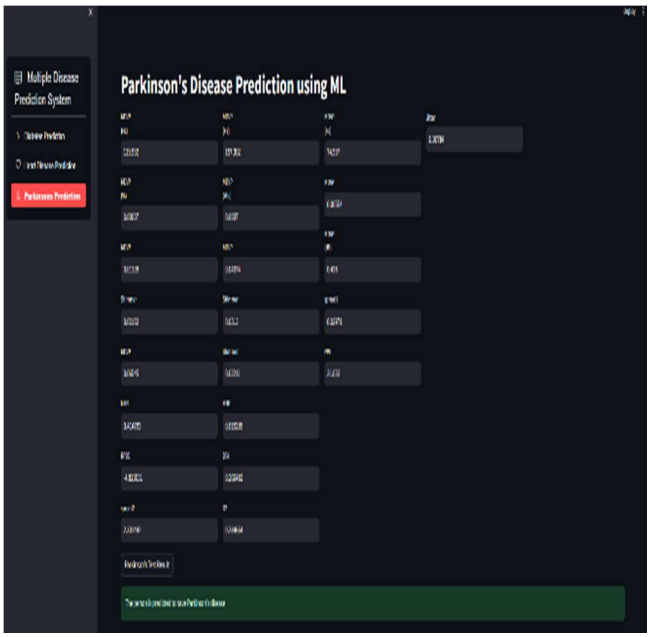

Figure.4. User interface for parkinson's disease prediction

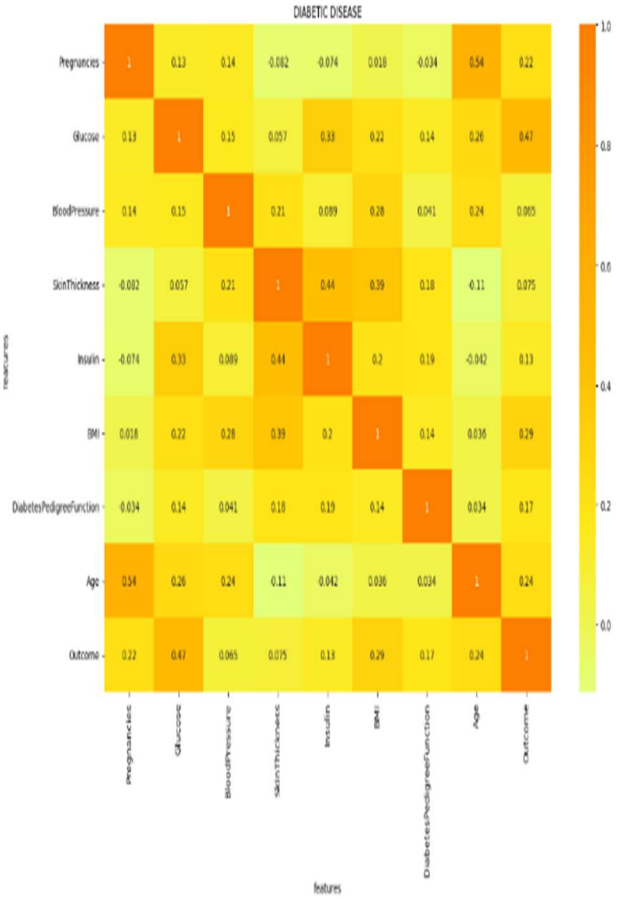

Figure 5. Heat Map for Diabetes Prediction

The above heatmap visually represents diabetic disease data, offering a graphical visualization that conveys patterns and trends within the dataset.

Figure 6. Heat Map for Heart Disease Prediction

The above heatmap provides a visual representation of the data pertaining to heart disease prediction, offering an insightful graphical overview that facilitates the observation of patterns and correlations within the dataset.

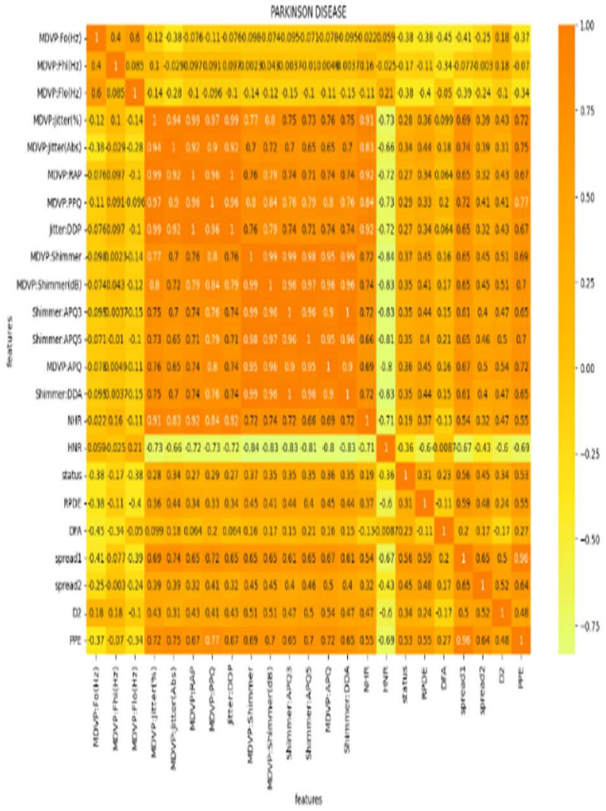

Figure 7. Heat Map for Parkison's Disease Prediction

The above heat map employed for predicting Parkinson's disease visually dis-plays the correlation matrix, illustrating the relationships among various parame-ters utilized in the analysis, such as tremor amplitude, bradykinesia, rigidity, and other relevant features.

V. Conclusion And Future Work

This paper represents a significant step forward in the realm of health prediction systems, offering a comprehensive and inclusive approach to multi-disease predic-tion. The integration of various machine learning algorithms has demonstrated promising results, showcasing the system's ability to predict diseases such as diabetes, heart disease, and Parkinson's disease with notable accuracy. As we envision the future, our focus remains on expanding the system's disease coverage, enhancing prediction accuracy, and improving user accessibility. The ultimate aim is to contribute to the early detection of diseases, providing individuals with timely information to facilitate proactive healthcare decisions. By continually refining and expanding the capabilities of our system, we aspire to make a meaningful impact on public health, reducing mortality rates through precise and accessible disease forecasts. The inclusion of features like a chatbot reflects our commitment to user-friendly interaction and support, further enhancing the overall user experi-ence. In essence, our project lays the foundation for a more informed and proactive approach to healthcare, aligning with the broader goal of improving health out-comes and promoting well-being.As we anticipate the future, there is considerable room for the advancement and enhancement of our system. Our strategy involves extending the API to encompass more diseases, thereby expanding its ability to predict and evaluate a wide array of health conditions.

REFERENCES

[1] Arumugam, K., et al., authored a paper titled "Predicting Multiple Diseases through Machine Learning Algorithms," published in Materials Today: Proceed-ings (Volume 80, 2023), encompassing insights into disease prediction using various machine learning techniques. The study delves into the anticipation of multiple ailments, presenting findings in the range of pages 3682-3685.

[2] Lu Men and colleagues authored a paper titled "Multi-disease prediction using LSTM recurrent neural networks," published in Expert Systems with Applica-tions, volume 177 in 2021, with the article number 114905.

[3] Xie, Shuxuan, Zengchen Yu, and Zhihan Lv authored a paper entitled "Survey on Multi-Disease Prediction Using Deep Learning," published in CMES-Computer Modeling in Engineering & Sciences (Volume 128, Issue 2, 2021).

[4] Harimoorthy, Karthikeyan, and Menakadevi Thangavelu authored a paper ti-tled "Multi-Disease Prediction Model Using Enhanced SVM-Radial Bias Tech-nique in Healthcare Monitoring System," published in the Journal of Ambient Intelligence and Humanized Computing (Volume 12, 2021, pages 3715-3723).

[5] Akkem Yaganteeswarudu authored a paper titled "Multi disease prediction model by using machine learning and Flask API," presented at the 2020 5th Inter-national Conference on Communication and Electronics Systems (ICCES), and published by IEEE.

[6] Sohyun Bang and collaborators authored a paper titled "Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data," published in Scientific Reports, volume 9, issue 1 in 2019, with the article number 10189.

[7] Anil Kumar Dubey authored a paper titled "Optimized hybrid learning for mul-ti disease prediction enabled by lion with butterfly optimization algorithm," pub-lished in Sādhanā, volume 46, issue 2 in 2021, page 63.

[8] Ralph B. D'Agostino and colleagues conducted a study titled "Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation," published in JAMA (Journal of the American Medical As-sociation), volume 286, issue 2 in 2001, pages 180-187.

[9] Marimuthu, M., M. Abinaya, K. S. Hariesh, K. Madhankumar, and V. Pavithra. "A review on heart disease prediction using machine learning and data analytics approach." International Journal of Computer Applications 181, no. 18 (2018): 20-25.

[10] Vasavi, D., et al. "MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING."

[11] Laxmi Deepthi Gopisetti and collaborators authored a paper titled "Multiple Disease Prediction System using Machine Learning and Streamlit," presented at the 2023 5th International Conference on Smart Systems and Inventive Technol-ogy (ICSSIT), and published by IEEE.

[12] Khadir, M. A., Mohd, A., Ali, M., & Khan, P. A. (2023). Multiple Disease Pre-diction System Using Machine Learning. Mathematical Statistician and Engineer-ing Applications, 72(1), 1435-1445.

[13] Kumari, V. Anuja, and R. Chitra et al,published a paper titled "Classification of diabetes disease using support vector machine" in the International Journal of Engineering Research and Applications in 2013, specifically in volume 3, issue 2, with pages 1797-1801.

[14] Jindal, Harshit, et al., conducted a study titled "Heart disease prediction using machine learning algorithms," which was

published in the IOP Conference Series: Materials Science and Engineering in 2021. The specific details include Volume 1022, Issue 1, by IOP Publishing.

[15] Nilanjan Dey and colleagues published a paper titled "Customized VGG19 architecture for pneumonia detection in chest X-rays" in Pattern Recognition Letters, volume 143 in 2021, pages 67-74.