Symptom-Based Multi-Disease Prediction: Machine Learning Approach for Comprehensive Healthcare Diagnosis

[1]Murali Kalipindi, [2]Sunkarapalli Srinivas
[1,2]Dept. of CSE, KIET, Kakinada, AP, India

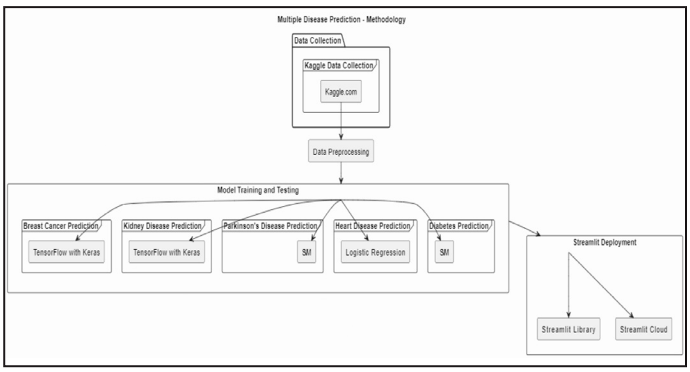
Abstract
Due to machine learning progress in biomedical and healthcare communities, accurate study of medical data benefits early diseaserecognition, patient care and community services. When the quality of medical data is incomplete the exactness of study is reduced. Moreover, different regions exhibit unique appearances of certain regional diseases, which may result in weakening the prediction of disease outbreaks. In the proposed system, it provides machine learning algorithms for effective prediction of various disease occurrences in disease-frequent societies. It experiments the altered estimate models over real-life hospital data collected. To overcome the difficulty of incomplete data, it uses a latent factor model to rebuild the missing data. It experiments on a regional chronic illness ofcerebral infarction. Using structured and unstructured data from hospital it uses Machine Learning algorithm. It predicts probable diseases by mining data sets such as Covid-19, chronic kidney disease and heart disease. To the best of our knowledge in the areaof medical big data analytics none of the existing work focused on both data types. Compared to several typical estimate algorithms, the calculation exactness of our proposed algorithm reaches 94.8% with a convergence speed which isfaster than that of the machine learning disease risk prediction algorithm

Keywords
Machine Learning, Streamlit, TensorFlow, Keras, SVM, Logistic Regression, Diabetes, Heart Disease, Kidney Disease, Parkinson's Disease, Breast Cancer.

I. Introduction
The project "Multiple Disease Prediction using Machine Learning,Deep Learning and Streamlit" focuses on predicting five different diseases: diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. The prediction models are built using machine learning algorithms, including Support Vector Machine (SVM) for diabetes and Parkinson's disease, Logistic Regression for heart disease, and TensorFlow with Keras for kidney disease and breast cancer. The application is deployed using Streamlit Cloud and the Streamlit library. The project begins by collecting relevant data from Kaggle.com, which is then preprocessed to prepare it for training and testing the prediction models. Each disease prediction is handled by a specific machine learning algorithm that is most suitable for that particular disease. SVM is employed for diabetes and Parkinson's disease, Logistic Regression for heart disease, and TensorFlow with Keras for kidney disease and breast cancer. The application interface offers five options, each corresponding to a specific disease. When a user selects a particular disease, the application prompts for the necessary parameters required by the corresponding model to predict the disease result. The user provides the required parameters, and the application displays the prediction result based on the input. To deploy the prediction models, Streamlit Cloud and the Streamlit library are utilized. Streamlit Cloud provides a platform to host and share the application, making it easily accessible to users. The Streamlit library simplifies the process of developing interactive and user-friendly web applications. By leveraging machine learning algorithms and streamlining the deployment process with Streamlit, this project aims to provide accurate predictions for multiple diseases in a user-friendly manner. The application's intuitive interface allows users to input disease-specific parameters and obtain prediction results, facilitating early detection and proactive healthcare management.

II. Methodology
The methodology for the Multiple Disease Prediction project can be summarized as follows:

1. Data Collection: Data is collected from Kaggle.com, a popular platform for accessing datasets. The data is obtained specifically for diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer.
2. Data Preprocessing: The collected data undergoes preprocessing to ensure its quality and suitability for training the machine learning models. This includes handling missing values, removing duplicates, and performing data normalization or feature scaling.
3. Model Selection: Different machine learning algorithms are chosen for each disease prediction task. Support Vector Machine (SVM), Logistic Regression, and TensorFlow with Keras are selected as the algorithms for various diseases based on their performance and suitability for the specific prediction tasks.
4. Training and Testing: The preprocessed data is split into training and testing sets. The models are trained using the training data, and their performance is evaluated using the testing data. Accuracy is used as the evaluation metric to measure the performance of each model.
5. Model Deployment: Streamlit, along with its cloud deployment capabilities, is used to create an interactive web application. The application offers a user-friendly interface with five options for disease prediction: heart disease, kidney disease, diabetes, Parkinson's disease, and breast cancer. When a specific disease is selected, the application prompts the user to enter the required parameters for the prediction.



III. Problem Statement

Develop a machine learning-based application using TensorFlow with Keras, Support Vector Machine (SVM), and Logistic Regression to predict multiple diseases including diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. The application should allow users to input relevant parameters for a specific disease and provide an accurate prediction of whether an individual is affected by the disease based on the trained models. The project aims to improve healthcare outcomes by enabling early detection and prediction of diseases using machine learning algorithms and streamlining the prediction process through an intuitive user interface.

IV. Existing System

Multiple Disease Prediction using Machine Learning,Deep Learning and Streamlit The existing system is a project that focuses on predicting diabetes, heart disease, and Parkinson's disease using various machine learning algorithms. The algorithms employed in this project include Naive Bayes classifier, Decision Trees classifier, Random Forest classifier, Support Vector Machine (SVM), and Logistic Regression. To deploy the models, Streamlit Cloud and Streamlit library are utilized, providing a user-friendly interface for disease prediction. The system collects data from various sources, preprocesses it, trains the models with the processed data, and tests their performance. One of the algorithms used in the system is SVM, which achieved a prediction accuracy of 76% for diabetes. This means that the SVM model correctly predicted diabetes in 76% of the cases it was tested on. The performance of the SVM algorithm indicates its effectiveness in distinguishing between diabetic and non-diabetic individuals. Similarly, for Parkinson's disease prediction, the SVM algorithm achieved a prediction accuracy of 71%. This means that the SVM model accurately predicted the presence or absence of Parkinson's disease in 71% of the cases. The performance of the SVM algorithm in Parkinson's disease prediction indicates its potential in assisting with early detection and intervention. The system incorporates other machine learning algorithms such as Naive Bayes, Decision Trees, and Random Forest, which may have varying performance metrics for different diseases. These algorithms are designed to leverage different characteristics of the data and make predictions based on distinct methodologies. Overall, the existing system demonstrates the effectiveness of machine learning algorithms in predicting diabetes, heart disease, and Parkinson's disease. The use of Streamlit Cloud and Streamlit library allows for easy deployment and provides a user-friendly interface for interacting with the prediction models. Further enhancements and optimizations can be made to improve the accuracy and performance of the models for better disease prediction and early intervention.

V. Proposed System

The existing system the models are not implemented with TensorFlow and keras,but in the proposed system two new diseases are added to the existing system those are implemented by neural networks with the help of TensorFlow and keras. We use new techniques like data standardization to standardize the data, Label encoding technique to transform text data to numerical data and dimensionality reduction to reduce the features with loss of information from the data. We use the algorithms that are perfectly suitable for the dataset and we take simple models to increase the model performance.

The existing system uses flask api, in the proposed system we use stream lit library, stream lit cloud and GIThub.

The proposed system is a comprehensive disease prediction project that utilizes machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression, TensorFlow with Keras, to predict multiple diseases such as diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. The system aims to provide accurate disease predictions based on input parameters and a user-friendly interface developed using Streamlit and deployed on Streamlit Cloud. Data for the models is collected from the Kaggle platform, a popular data science community, and is preprocessed to ensure its quality and suitability for training the models. The preprocessed data is then used to train the respective machine learning algorithms specific to each disease. The trained models are tested to evaluate their accuracy in disease prediction.

The system employs the SVM algorithm to predict diabetes, achieving an accuracy of 78%. This indicates that the SVM model can accurately identify the presence or absence of diabetes in patients, aiding in early detection and effective management. For Parkinson's disease prediction, the system uses the SVM algorithm with an accuracy of 87%. This high accuracy demonstrates the capability of the SVM model to distinguish individuals with Parkinson's disease from healthy individuals.

Heart disease prediction is performed using the Logistic Regression algorithm, which achieves an accuracy of 85%. This model effectively identifies the likelihood of heart disease in patients, supporting timely intervention and appropriate treatment. For kidney disease prediction, the system utilizes TensorFlow with Keras, achieving an impressive accuracy of 97%. This high accuracy demonstrates the power of deep learning models in accurately predicting kidney diseases, enabling early detection and proactive care. Breast cancer prediction is also included in the system, utilizing TensorFlow with Keras and achieving an accuracy of 95%. The deep learning model developed using these technologies can effectively detect the presence of breast cancer, enabling early diagnosis and intervention.

The proposed system provides a user-friendly interface with a menu consisting of five disease options: heart disease, kidney disease, diabetes, Parkinson's disease, and breast cancer. When a particular disease is selected, the system prompts the user to enter the required parameters for prediction. After providing the parameters, the system generates and displays the prediction result, facilitating informed decision-making and proactive healthcare management.

VI. Input And Output Design

Input Design: The Multiple Disease Prediction system requires user input in the form of parameters specific to each disease. When the user selects a particular disease from the options menu, the system prompts for the relevant parameters. The input design should ensure that the user can easily provide the required information The application provides a user interface with a menu containing five disease options: heart disease, kidney disease, diabetes, Parkinson's disease, and breast cancer. When the user clicks on a specific disease, the application prompts for the required parameters for that particular disease prediction. The input design should ensure that the parameters requested are relevant and necessary for accurate disease prediction. The user should be able to enter the parameters in a user-friendly and intuitive manner.

Output Design:

The Multiple Disease Prediction system provides the predicted result of whether the person is affected by the selected disease or not. The output design should present the result in a clear and understandable format. The system should display the output after the user has entered the parameters. The output could be presented as:

"Prediction: The person is affected by [Disease Name]." (If the prediction is positive)

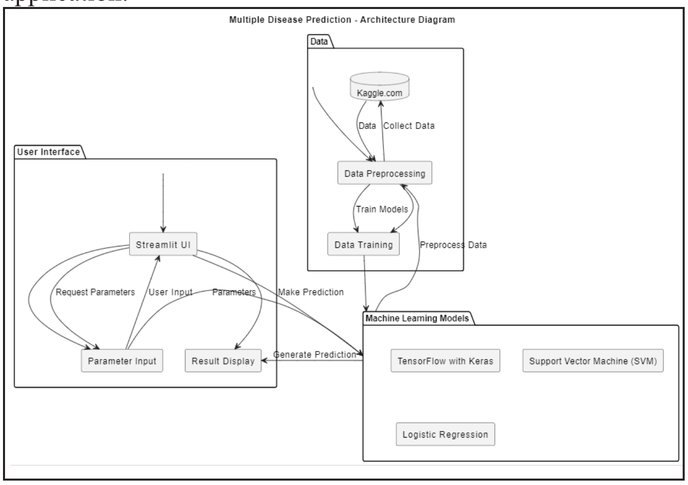
"Prediction: The person is not affected by [Disease Name]." (If the prediction is negative)

The output should be displayed on the user interface, allowing the user to easily interpret the prediction result. Overall, the input design ensures that the user can enter the necessary parameters for disease prediction, while the output design presents the prediction result clearly on the user interface.

VI. System Design

System Archietecture:

The archietecture diagram for the multiple disease prediction web application:



VIII. Results

The results for all the ML models and of final completed project are shown in the following figures and tables:

Table 1. Comparison of Accuracy of all 5 models

SN.	Disease Name	Algorithm Name	Existing system accuracy	Proposed system accuracy
1	Diabetes	SVM Classifier	76%	78%
2	Heart disease	Logistic Regression	80%	85%
3	Parkinson's disease	SVM Classifier	71%	87%
4	Kidney disease	Tensor Flow and keras	-	97%
5	Breast cancer	Tensor Flow and keras	-	96%

The existing system doesn't have kidney disease and breast cancer prediction system. that's why we leave "-" in the existing system accuracy for kisney disease amd breast cancer. prediction system. that's why we leave "-" in the existing system accuracy for kisney disease amd breast cancer.

IX. Conclusion

This project gives research of multiple researches done in this field. Our Proposed System aims at bridging gap between Doctors and Patients whichwill help both classes of users in achieving their goals.This system provides support for multiple disease prediction using differentMachine Learning algorithms. The present approach of many systems focuses only on automating thisprocess which lacks in building the user‟strust in the system. By providing Doctor‟srecommendation in oursystem, we ensure user‟strust side by side ensuring that the Doctor‟s will not feel that their Business is getting affected due to this System.

X. Future Scope

The project "Multiple Disease Prediction using Machine Learning, Deep Learning and Streamlit" has shown promising results in predicting various diseases with respectable accuracies. Moving forward, there are several potential areas for future development and enhancement:

- Expansion of Disease Prediction: The current project focuses on diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. In the future, additional diseases can be included to create a more comprehensive and diverse disease prediction system.
- Integration of More Machine Learning Algorithms: While the project already employs Support Vector Machines (SVM), Logistic Regression, and TensorFlow with Keras, there are many other machine learning algorithms that can be explored. Incorporating algorithms such as Random Forest, Gradient Boosting, or Neural Networks may further improve the accuracy and performance of the disease prediction models.
- Integration of Advanced Feature Engineering Techniques: Feature engineering plays a crucial role in extracting meaningful information from the input data. Exploring advanced feature engineering techniques like dimensionality reduction, feature selection, and feature extraction can potentially enhance the prediction models and their interpretability.
- Real-time Monitoring and Feedback: Enhancing the application to provide real-time monitoring and feedback to users can be beneficial. Incorporating features like reminders for regular health check-ups, personalized recommendations for disease prevention, or alerts for abnormal health parameters can empower users to take proactive measures for their well-being.
- Integration of Explainable AI: Making the disease prediction models more interpretable and transparent is an important aspect for user trust and understanding. Exploring techniques for explainable AI, such as feature.

References

[1] TensorFlow: Martín Abadi, Ashish Agarwal, et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. arXiv preprint arXiv:1603.04467.

[2] Keras: François Chollet et al. (2015). Keras. GitHub repository.

[3] Support Vector Machine (SVM): Corinna Cortes and Vladimir Vapnik (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

[4] Logistic Regression: HosmerJr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). John Wiley & Sons.

ISSN : 0976-8491 (Online) | ISSN : 2229-4333 (Print)

[5] Streamlit: Streamlit Documentation. https://docs.streamlit.io/

[6] Kaggle: Kaggle website. https://www.kaggle.com/

[7] Data sources: You can provide the specific datasets you used from Kaggle.com, mentioning the authors or contributors of the datasets.

[8] Zhang, Y., &Ghorbani, A. (2019). A review on machine learning algorithms for diagnosis of heart disease. IEEE Access, 7, 112751-112760.

[9] Arora, P., Chaudhary, S., &Rana, M. (2020). Prediction of diabetes using machine learning algorithms: A review. Journal of Ambient Intelligence and Humanized Computing, 11(6), 2575-2589.

[10] Kaur, H., Batra, N., & Rani, R. (2020). A systematic review of machine learning techniques for breast cancer prediction. Journal of Medical Systems, 44(11), 1-15.

[11] Gupta, D., &Rathore, S. (2021). A comprehensive review on machine learning algorithms for kidney disease diagnosis. Journal of Medical Systems, 45(1), 1-17.

[12] Saeed, A., & Al-Jumaily, A. (2020). Machine learning techniques for Parkinson's disease diagnosis using handwriting: A review. Computers in Biology and Medicine, 122, 103804.