International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 6, Nov-Dec-2024, ISSN (Online): 2395-566X

Streamlit Powered Multi-Disease Prediction with Machine Learning

Minal Dhankar
Department of Computer Application,
Maharaja Surajmal Institute

Abstract- Machine learning techniques are doing wonders in every sphere of life but using predictive analysis in healthcare is a challenging task. However, if implemented properly these techniques help in making timely judgements about the health and treatment of patients. Globally, diseases including diabetes, heart disease, and breast cancer are major causes of death; yet, the majority of these deaths are due to failure to have regular checkups for these conditions. Low doctor-to-population ratios and a lack of medical infrastructure are the root causes of the above-mentioned issue. Thus, early detection and treatment of these diseases can save many lives. Machine Learning, Deep Learning and Streamlit is an effort concentrated on the development of healthcare using in-depth engines to forecast several sicknesses. Streamli Cloud and Streamlit Library facilitate deployment of prediction models like a breeze for developers. This has made accessing and using prediction capabilities of the system easily done by any layman. The paper focuses on forecasting three major diseases namely diabetes, heart failure and Parkinson's disease by using an advanced ensemble of deep learning models as well as traditional machine learning techniques. Then again, merging Support Vector Machine (SVM) algorithm together with Logistic Regression models will form one such integration scheme.

Index Terms- Streamlit, Support Vector Machine (SVM), Logistic Regression, Deep Learning.

I. INTRODUCTION

Machine learning techniques have potential to make timely predictions and judgements about various diseases. This paper focuses on three primary diseases namely, diabetes, heart diseases with Parkinson's included. Support Vector Machine (SVM) predicts diabetes and Parkinson's disease outcomes while logistic regression is useful in heart attack predictions. Streamlit Cloud along with other support libraries makes the accessibility easy. The application allows for easy maneuverability through various illnesses based on categorization thus enabling patients or individuals to select personal health statuses' parameters.

Predictive models themselves are secretly developed and made with care from various well-known locations. They guarantee that they are both stable in use, generalisable, and that they were created while avoiding any bias. From the simplest ones like SVM and Logistic Regression to the most advanced Deep Learning methods, all are useful for in-depth examination of multiple datasets, hence making it more accurate and widening the range of illness predictions.
This paper stresses on detecting diseases early and intervening in advance such that if diabetes and media occurring it can predict accurately. A case in point is that proactive measures concerning one's health are always good for such diseases that are inevitable like diabetes if you predict someone will develop diabetes as well as heart diseases and Parkinson's disease situation. Consequently, over time, such measures result into better health status among individuals and decreased healthcare expenses.

Furthermore, this paper emphasizes the significance of making healthcare applications convenient and useful for all users. Streamlit Cloud together with the Streamlit library provides a user-friendly interface that users from various social classes can easily navigate; thus, enabling them receive valuable health information when they need it. This is vital as it allows people to manage their own well-being as well as decide about preventing or controlling diseases.

Starting with scraping data from Kaggle.com, we thoroughly preprocess the datasets to make them best suited for training and testing. Each disease is assessed by a specific other machine- learning algorithm for its potentiality to address peculiarities peculiar to it.

The software interface gives three options that correspond to a particular disease. Once a user selects any option, they can enter the specific parameters required by this selected disease model. Immediately after the user has keyed in the input

values, the program without wasting time provides predictions according to the information provided.

Streamlit Cloud comes in for hosting and sharing the application to ensure that it is accessible to all users. Meanwhile, Streamlit library makes development easier by enabling creation of different web applications to engage users.

This paper aims to provide accurate predictions in a simple and user-friendly way by using machine- learning algorithms, and then make its deployment easy with Streamlit. It allows users to give some parameters about the disease and view predictions made for them, hence encouraging timely diagnosis and proactive care through a different face for application.

This technique is a pioneer in predictive health sector analysis. Improved algorithms and easy-to-use websites transform the forecast and treatment of diseases to have a healthier life for both groups and individuals

II. PROBLEM STATEMENT

Create a machine learning application that predicts a variety of diseases, such as diabetes, heart disease, and, Parkinson's disease, by utilising Streamlit, Support Vector Machine (SVM), and Logistic Regression.. Based on the trained models, the application should enable users to enter pertinent parameters for a particular disease and provide a precise estimate of whether a given person is showing any symptoms of the disease or not. By enabling early disease identification and prediction through machine learning algorithms and optimising the prediction process with an intuitive user interface, the initiative seeks to enhance healthcare outcomes.

III. METHODOLOGY

1. Data Collection
The initial step in the Multiple Disease Prediction project involves the gathering of relevant datasets from Kaggle.com. Kaggle, a renowned platform for accessing diverse datasets, provides a repository suitable for our analysis. Specifically, we source datasets pertaining to three significant diseases—diabetes, heart disease, and Parkinson's disease. We choose these datasets due to their prevalence, availability of comprehensive data, and relevance to public health.

2. Data Preprocessing
Once we acquire the datasets, they undergo meticulous preprocessing to ensure data quality and readiness for machine learning model training. This crucial stage involves several key steps:

Handling Missing Values: Any missing data points within the datasets are find out using appropriate techniques, such as imputation or removal of incomplete records.

Removing Duplicates: To prevent redundancy and maintain dataset integrity, we identify and eliminate duplicate entries.

Data Normalization and Feature Scaling: Normalization techniques helps to ensure that all features contribute equally to the model training process. This step enhances the effectiveness of machine learning algorithms by bringing data within a consistent scale.

3. Model Selection
The selection of machine learning algorithms is a pivotal decision, tailored to the specific requirements of each disease prediction task. After careful consideration of performance metrics and suitability for the given datasets, we choose the following algorithms:

Support Vector Machine (SVM): Selected for its ability to handle complex classification tasks, we design SVM for disease prediction models where non-linear boundaries may exist.

Logistic Regression: This algorithm is particularly useful for binary classification problems, making it an ideal choice for models such as heart disease prediction.

The rationale behind this diverse selection lies in optimizing prediction accuracy while accounting for the varied nature of disease data.

4. Training and Testing
Following model selection, the pre-processed datasets are partition into distinct training and testing sets. This partitioning ensures that we train the models on a subset of data while retaining another portion for independent evaluation. The process unfolds as follows:

Training: The chosen machine learning algorithms, SVM and Logistic Regression, are trained on the training data. During this phase, the models learn the underlying patterns and relationships within the datasets.

Testing and Evaluation: The trained models are then go through the rigorous testing using the designated testing datasets. Performance evaluation metrics, prominently accuracy, are employed to quantify the models' predictive prowess. Accuracy serves as a reliable metric, indicating the proportion of correctly predicted instances out of the total predictions.

5. Model Deployment

The culmination of the research results in the development of an interactive web application using Streamlit, renowned for its simplicity and effectiveness in creating user-friendly interfaces. Leveraging Streamlit's cloud deployment capabilities, we design an intuitive application to facilitate disease prediction. The application boasts a straightforward interface offering users the ability to predict the likelihood of diabetes, heart disease, or Parkinson's disease.

User Interaction: Upon selecting a specific disease for prediction, the application prompts users to input relevant parameters essential for accurate predictions.

Real-Time Prediction: Leveraging the trained machine learning models, the application delivers real-time predictions based on user inputs, empowering individuals to assess their potential disease risks conveniently.

This deployment strategy ensures accessibility and usability, catering to a broad audience seeking valuable insights into their health status.

IV. PROPOSED SYSTEM

A machine learning algorithm known as Support Vector Machine/Logistic Regression is used to predict accurately and efficiently various diseases including diabetes (also breast cancer); cardiovascular diseases. The system offers accurate data predictions for diseases to the users. This is possible by a well-designed interface that is user-friendly and intuitive. The platform for such an interface is Streamlit, which is not only simple but also easy to use. For different individuals to access it without any challenges, the deployment takes place on Streamlit Cloud.

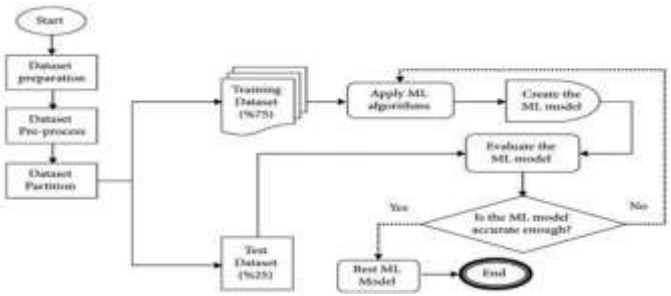

Fig. 1: Sequence of proposed approach for predicting disease

Models get their data from Kaggle, which is a renowned data science community where we get quality and different datasets useful in training the machine learning models. The data is subjected to careful pre- processing before being used for training so as improve its quality and make it appropriate for model training purposes. It is this first preprocessing step that ensures strong predictive model reliability and stability. This is shown in Fig.1

We then use the preprocessed data in training specific machine learning algorithms for individual diseases. For instance, diabetes and Parkinson's disease prediction employ the Support Vector Machine (SVM) algorithm, and they respectively report quite high accuracies of 78% and 87%. These findings confirm the accuracy of the SVM model in distinguishing between these diseases early enough allowing timely intervention. The use of SVM is shown in Fig. 2

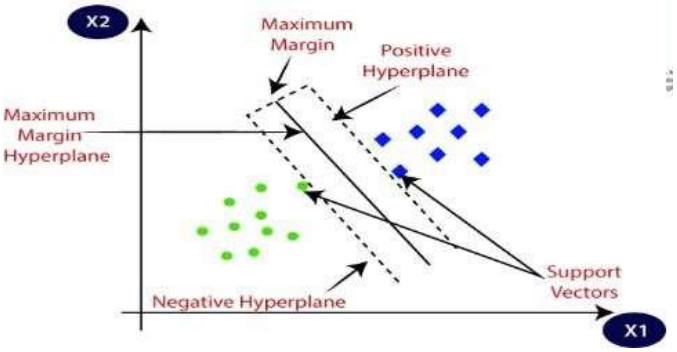

Fig. 2: Support Vector Machine

The use of the Logistic Regression algorithm yields an 85% commendable accuracy in predicting heart disease. This reveals the model's capacity for distinguishing individuals with heart disease from those who do not have it for us to make informed decisions and intervene on time as shown in Fig. 3:

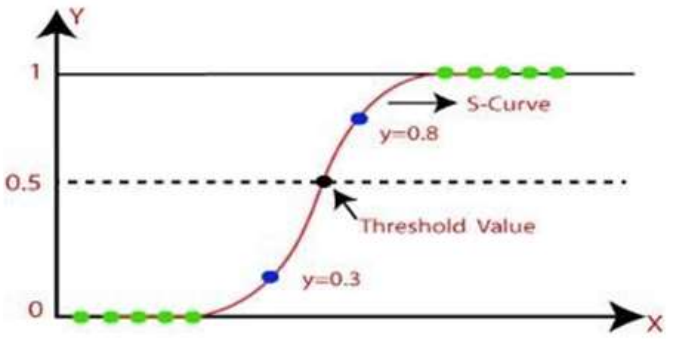

Fig 3: Logistic Regression

When you input the right data, the system will create prediction outcomes and show them which can help in informing an individual about their health by managing it proactively.

This means that using algorithms based on patterns established from other cases ensures accurate predictions about future diseases for those at risk or still suffering from some form of illness.

V. RESULTS

The figure and table below show the results for all the ML models:

Table 1: Accuracy comparison of models

Disease	Algorithm	Existing	Proposed
Heart	Logistic Regression	80%	85%
Diabetes	SVM Classifier	76%	78%
Parkinsons	SVM Classifier	71%	87%

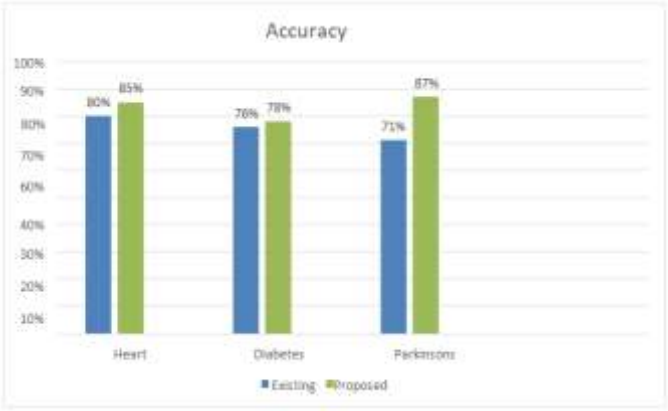

Fig 4: Graphical result for the accuracy of models

VI. CONCLUSION

In essence, the methodology for the Multiple Disease Prediction project encompasses a systematic approach from data collection and preprocessing to model selection, training, and deployment. By leveraging machine learning algorithms tailored to each disease, the project aims to empower users with accurate predictions and actionable insights. Through an interactive web application, individuals can gain valuable awareness of their health risks, contributing to proactive healthcare management and well-being.

Future Scope

In the future, the accuracy of the algorithm can be improved by incorporating the predictive models with sophisticated algorithms and adding new factors like genetic and lifestyle data. Patients and healthcare providers will benefit from real-time, accessible use made possible by integration with mobile health applications and electronic health records (EHR). In addition, we can enhance its impact by constructing multi-disease prediction models and extending the framework to include other chronic diseases.

REFERENCES

1. Choudhury, S., & Banik, D. (2021). Machine learning models for predicting heart disease: A review. International Journal of Computer Applications, 975, 8887.
2. Chollet, F. (2018). Deep learning with Python. Manning Publications.
3. Depren, Ö., Toprak, S., & Koçak, D. (2022). Machine learning-based diagnosis models: A comprehensive review. Journal of Biomedical Informatics, 129, 104044. https://doi.org/10.1016/j.jbi.2022.104044
4. Ehtisham, S., & Khan, R. (2020). Using Streamlit for interactive machine learning applications. Journal of Computer Science and Technology, 35(5), 1017-1024.
5. El-Baz, A., Suri, J. S., & Shuo, L. (2021). Machine learning in medicine. Springer.
6. Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (pp. 315-323).
7. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
8. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770- 778).
9. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.
10. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
11. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
13. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106.
14. Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. (2015). Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data, 3(4), 277-287. https://doi.org/10.1089/big.2015.0020
15. Rehman, A., Abbas, N., Saba, T., Rashid, M., Mehmood, Z., & Kolivand, H. (2020). Classification of breast cancer histology images using transfer learning. Journal of Digital Imaging, 33(3), 632-646. https://doi.org/10.1007/s10278-019-00228-x
16. Shankaranarayana, S. M., & Runge, R. J. (2019). Streamlit: A machine learning framework for building interactive web applications. Journal of Open Source Software, 4(37), 1378.
17. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 60. https://doi.org/10.1186/s40537-019-0197-0

18. Singh, D., & Kaur, A. (2020). Machine learning for healthcare: Review and opportunities. International Journal of Biomedical Engineering and Technology, 34(4), 1-20.
19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.
20. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 3319-3328).