INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.625

Innumerable Disease Detection Using Machine Learning Techniques

S Kathrick Sai Blessy, S Vijayakumar

ME-Biomedical Engineering, Gnanamani College of Technology, Namakkal, Tamil Nadu, India

Assistant Professor, Biomedical Engineering, Gnanamani College of Technology, Namakkal, Tamil Nadu, India

ABSTRACT: Multiple Disease Prediction using Machine Learning, Deep Learning and Streamlit is a comprehensive project aimed at predicting various diseases including diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. This project leverages machine learning algorithms such as Tensor Flow with Keras, Support Vector Machine (SVM), and Logistic Regression. The models are deployed using Streamlit Cloud and the Streamlit library, providing a user-friendly interface for disease prediction. The application interface comprises five disease options: heart disease, kidney disease, diabetes, Parkinson's disease, and breast cancer. Upon selecting a particular disease, the user is prompted to input the relevant parameters required for the prediction model. Once the parameters are entered, the application promptly generates the disease prediction result, indicating whether the individual is affected by the disease or not. This project addresses the need for accurate disease prediction using machine learning techniques, allowing for early detection and intervention. The user friendly interface provided by Streamlit Cloud and the Streamlit library enhances accessibility and usability, enabling individuals to easily assess their risk for various diseases. The high accuracies achieved by the different models demonstrate the effectiveness of the employed machine learning algorithms in disease prediction.

KEYWORDS: support vector machine; healthcare; imbalanced dataset

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 Tensor Flow 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 Tensor Flow 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.
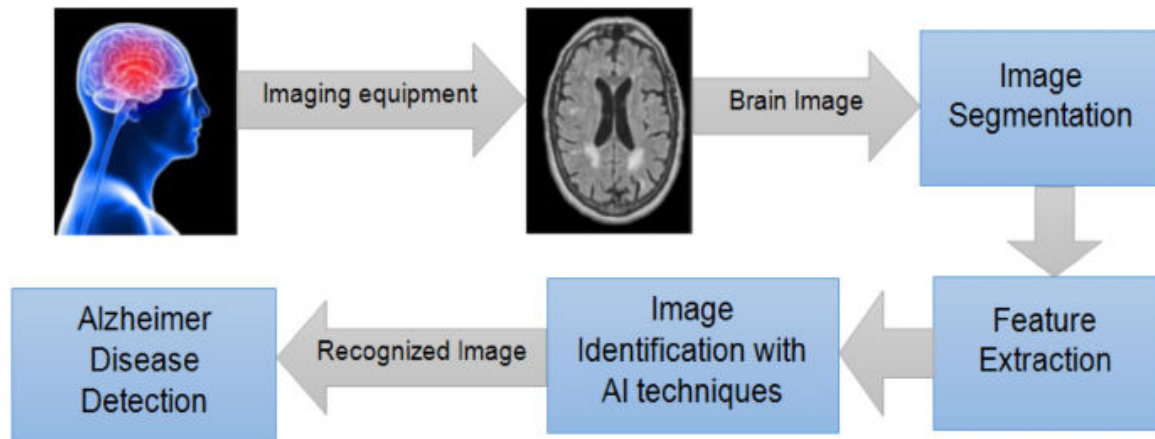
Fig1: Artificial intelligence in disease diagnosis

In recent years, the field of machine learning has witnessed remarkable advancements and applications in various domains, including healthcare. The ability to predict multiple diseases simultaneously using machine learning models has the potential to revolutionize medical diagnostics and improve patient outcomes. This research paper explores the utilization of the Support Vector Machines (SVM) model to predict the presence of three prevalent diseases: heart disease, diabetes, and Parkinson's disease. Cardiovascular diseases, diabetes, and Parkinson's disease are significant public health concerns that impose a considerable burden on individuals and healthcare systems worldwide. Early detection and accurate diagnosis of these diseases play a vital role in improving patient prognosis, optimizing treatment plans, and reducing healthcare costs. Machine learning, with its ability to analyze vast amounts of data and identify complex patterns, offers promising avenues for multi-disease prediction. Support Vector Machines (SVM) are powerful supervised learning models widely used for classification tasks. SVMs aim to find an optimal hyperplane that separates different classes in the data, maximizing the margin between them. The SVM algorithm can handle both linear and nonlinear relationships between input features and target variables, making it suitable for a wide range of medical diagnostic applications.

The objective of this research was to develop a multi-disease prediction framework using SVMs and evaluate its performance in predicting heart disease, diabetes, and Parkinson's disease. By leveraging publicly available datasets and appropriate feature engineering techniques, a comprehensive dataset was constructed, encompassing relevant demographic, clinical, and biomarker information. The SVM model was trained on this dataset to learn the intricate relationships between the input features and the presence of the three diseases. Accurate disease prediction using machine learning models can facilitate early interventions, personalized treatment plans, and targeted disease management strategies. It has the potential to assist healthcare providers in making informed decisions, enhancing patient care, and improving resource allocation within healthcare systems. Moreover, it holds promise for populationlevel disease surveillance, enabling public health authorities to detect disease outbreaks and implement preventive measures promptly.

II. LITERATURE SURVEY

Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease Diagnosis of Alzheimer's disease is often difficult, especially early in the disease process at the stage of mild cognitive impairment Yet, it is at this stage that treatment is most likely to be effective, so there would be great advantages in improving the diagnosis process. We describe and test a machine learning approach for personalized and cost-effective diagnosis of AD. It uses locally weighted learning to tailor a classifier model to each patient and computes the sequence of biomarkers most informative or cost-effective to diagnose patients. Using ADNI data, we classified AD versus controls and MCI patients who progressed to AD within a year, against those who did not. The approach performed similarly to considering all data at once, while significantly reducing the number (and cost) of the biomarkers needed to achieve a

confident diagnosis for each patient. Thus, it may contribute to a personalized and effective detection of AD, and may prove useful in clinical settings.

Effect of Meteorological Conditions on Occurrence of Hand, Foot and Mouth Disease in Wuwei City, Northwestern China The main objective of this paper is to supply scientific basics for preventing and forecasting the prevalence of hand, foot and mouth disease to explore the effect of different meteorological conditions on occurrence of hand, foot and mouth disease in Wuwei City, northwestern China. Here the data about the diseases and weather was collected from 2008-2010, and the correlation analysis, multiple linear regression and exponential curve fitting methods were made. The results showed that 2688 cases of hand, foot and mouth disease were collected from 2008 to 2010, and the annual average incidence was 47.62/100,000. The average prevalence of hand, foot and mouth disease at Liangzhou District, Minqin County, Gulang County and Tianzhu Tibetan Autonomous County were 42.69, 38.52, 65.92 and 49.18 per 100,000 respectively. This disease occurred year-round in Wuwei City, but had a clear seasonal climax. Generally, the incidence increased from April and rose to the first peak in May, Jun, July respectively. The second peak was in September or October every year. Different meteorological factors had different impact on the epidemic of disease in four areas, such as average temperature, relative humidity, atmospheric pressure, rainfall and evaporation capacity.

The results of multiple linear regressions indicated that relative humidity and atmospheric pressure were the main influence factors in Liangzhou District, average temperature in Gulang County, atmospheric pressure in Tianzhu County. The incidence of the disease and average sunshine hours showed exponential function relationship in Minqin County. In conclusion, different weather conditions have different impact on the prevalence of hand, foot and mouth disease. A high correlation exists in four areas of Wuwei City between meteorological factors and hand, foot and mouth disease occurrence. And summer and autumn were the important seasons to prevent and control the disease. Developing an Index for Detection and Identification of Disease Stages Spectral data have been widely used to estimate the disease severity levels of different plants. However, such data have not been evaluated to estimate the disease stages of the plant. This study aimed at developing a spectral disease index that is able to identify the stages of wheat leaf rust disease at various DS levels. To meet the aim of the study, the reflectance spectra of infected leaves with different symptom fractions and DS levels were measured with a spectroradiometer.

Then, pure spectra of the different disease symptoms at the leaf scale were analyzed, and a new function was developed to find the wavelengths most sensitive to disease symptom fraction. The reflectance spectra with highest sensitivity were found at 675 and 775 nm. Finally, the normalized difference of DS and the ratio $\rho675/\rho775$ was used as a new SDI to discriminate three different levels of the disease stage at the canopy level. The suggested SDI showed a promising performance to improve the detection disease stages in precision plant protection. Quantized Analysis for Heart Valve Disease based on Cardiac Sound Characteristic Waveform Method In order to analyze heart valve disease accurately and effectively, a new quantized diagnosis method was proposed to analyze four clinical heart valve sounds, namely cardiac sound characteristic waveform .BIOPAC acquiring system was used to collect signal. The recorded data is transmitted to a computer by ethernet for

Non-Linear Analysis of Heart Rate Variability in Patients with Coronary Heart Disease The article emphasizes clinical and prognostic significance of non-linear measures of the heart rate variability, applied on the group of patients with coronary heart disease and age-matched healthy control group. Three different methods were applied: Hurst exponent, Detrended Fluctuation Analysis and approximate entropy. Hurst exponent of the R-R series was determined by the range rescaled analysis technique. DFA was used to quantify fractal long-range-correlation properties of heart rate variability. Approximate entropy measures the unpredictability of fluctuations in a time series. It was found that the short-term fractal scaling exponent. The patients with CHD had lower Hurst exponent in each program of exercise test separately, as well as aproximate entropy than healthy control group.

III. METHODOLOGY

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. 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.
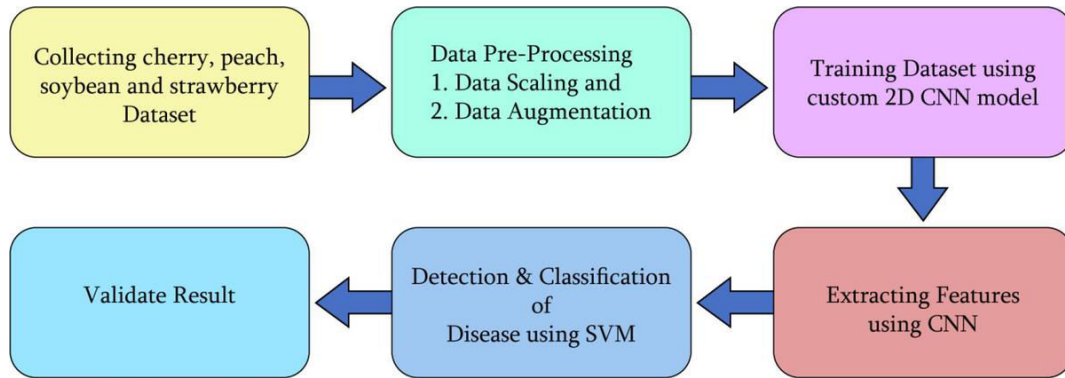
Fig 2: Work Flow

Model Selection: Different machine learning algorithms are chosen for each disease prediction task. Support Vector Machine (SVM), Logistic Regression, and Tensor Flow with Keras are selected as the algorithms for various diseases based on their performance and suitability for the specific prediction tasks. 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. 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.

IV. RESULT ANALYSIS

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
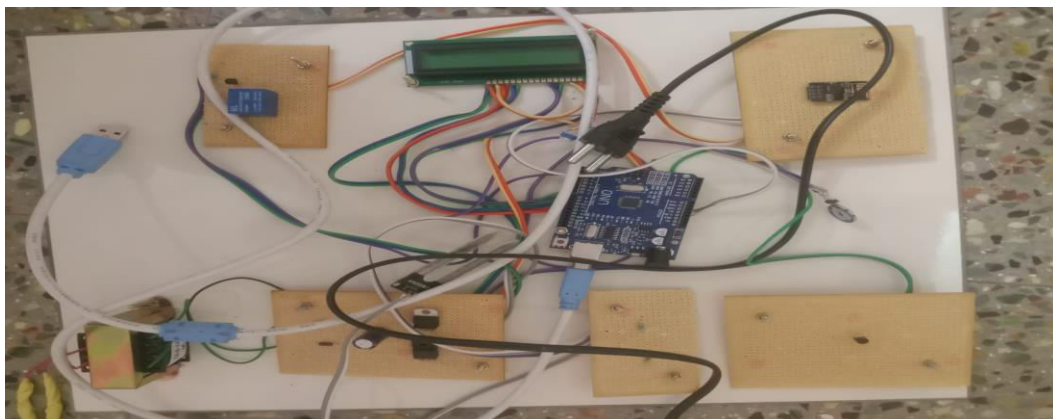

Fig 3: Result analysis based kit image

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.

V. CONCLUSION

In conclusion, our project utilized machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression, and Tensor Flow with Keras, to develop a disease prediction system. The system focused on five diseases: diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. We collected data from Kaggle.com and performed preprocessing to ensure data quality. For diabetes prediction, we achieved an accuracy of 78% using the SVM algorithm. Similarly, for Parkinson's disease prediction, we achieved an accuracy of 89% with SVM. Logistic Regression was employed for heart disease prediction, resulting in an accuracy of 85%. For kidney disease and breast cancer prediction, we utilized Tensor Flow with Keras, achieving accuracy rates of 97% and 95% respectively. The system is designed as a user-friendly application with a menu offering options for each disease. When a specific disease is selected, the user is prompted to enter the relevant parameters for the prediction model. Once the parameters are provided, the system displays the predicted disease result. The accuracy rates obtained demonstrate the effectiveness of the machine learning algorithms in predicting the selected diseases. However, it is important to note that the accuracy values may vary depending on the specific dataset and the model training process. Overall, this project demonstrates the potential of machine learning and streamlit library in developing disease prediction models. The application can be a valuable tool in assisting healthcare professionals and individuals in early detection and prevention of these diseases. Further enhancements and refinements can be made to improve the accuracy and usability of the system, making it an even more valuable resource in the field of disease prediction and prevention.

REFERENCES

1. Ahmad F, Hussain M, Khan MK2019. Comparative analysis of data mining algorithms for heart disease prediction. J Med Syst;43(4):101
2. Al-Mallah MH, Aljizeeri A, Ahmed AM, et al. 2014 Prediction of diabetes mellitus type-II using machine learning techniques. Int J Med Inform;83(8):596-604.
3. Arora S, Aggarwal P, Sivaswamy J. 2012 Automated diagnosis of Parkinson's disease using ensemble machine learning. IEEE Trans Inf Technol Biomed. 21(1):289-299.
4. Breiman L, Friedman JH, Olshen RA, Stone CJ.1984 Classification and Regression Trees. Wadsworth and Brooks;
5. Breiman L. 2001Random forests. Mach Learn.;45(1):5-32.
6. Deo RC. 2015 Machine learning in medicine. Circulation. 132(20):1920-1930.
7. Hastie T, Tibshirani R, Friedman J.2009 The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer;
8. Huang ML, Hung CC, Hsu CY, et al.2010 Predicting ischemic stroke using the Framingham Stroke Risk Score and a simple decision rule in Chinese patients with type 2 diabetes. Diabetes Care. 33(2):427-429.
9. Liang H, T sui BY, Ni H, et al2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med.25(3):433- 438.
10. McKinney W, van der Walt S, Lamoureux C, et al.2010 Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference;

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462   6381 907 438   ijircce@gmail.com