Navigating Health Data: Building Predictive Models for Disease diagnosis With Machine Learning

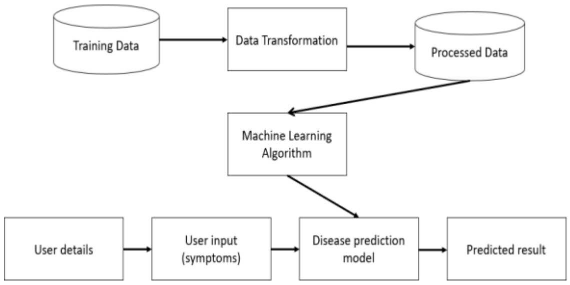
M. Sowmya[1], P. Shravya[2], G. Durga Manikanta Satya Srinivas[3], M. NIkitha[4], Mr.G.Mukesh[5], Mr.G.Rakesh Reddy [6]

[1,2,3,4]Students, Department of CSE(IOT), Sphoorthy Engineering College, Hyderabad, India.
[5]Assistant Professor, Department of CSE(IOT), Sphoorthy Engineering College, Hyderabad, India.
[6]Professor & HOD, Department of CSE(IOT), Sphoorthy Engineering College, Hyderabad, India.

***Abstract*—The dependency on computer-based technology has resulted in storage of lot of electronic data in the health care industry. As a result of which, health professionals and doctors are dealing with demanding situations to research signs and symptoms correctly and perceive illnesses at an early stage. However, Machine Learning technology have been proven beneficial in giving an immeasurable platform in the medical field so that health care issues can be resolved effortlessly and expeditiously. Disease Prediction is a Machine Learning based system which primarily works according to the symptoms given by a user. The disease is predicted using algorithms and comparison of the datasets with the symptoms provided by the user. To the best of our knowledge in the area of 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 is faster than that of the machine learning disease risk prediction algorithm.**

I. INTRODUCTION

The Earth is going through a purplish patch of technology where the demand of intelligence and accuracy is increasing behind it. Today's people are likely addicted to internet but they are not concerned about their physical health. People ignore the small problem and don't to visit hospital which turn into serious disease with time. Taking the advantage of this growing technology, our basis aim is to develop such a system that will predict the multiple diseases in accordance with symptoms put down by the patients without visiting the hospitals physicians. Machine Learning is a subset of AI that is mainly deal with the study of algorithms which improve with the use of data and experience. Machine Learning has two phases i.e. Training and Testing. Machine Learning provides an efficient platform in medical field to solve various healthcare issues at a much faster rate. As the importance of early detection of CVD is increasingly being realized, there is a definite need of developing ML-based prediction system for CVDs specifically suitable for Indian scenario. This study was carried out with the following objectives: a) Development of a high-performance and cost-effective ML-based heart disease prediction system using routine clinical data specifically suited for Indian population and b) Deployment of the prediction system in public cloud to ensure easy accessibility via Internet particularly beneficial for rural areas in India.



II. RELATED WORK

Numerous studies have explored the use of machine learning for disease detection and prediction across different domains of healthcare. In the context of Alzheimer’s Disease, Escudero et al. applied locally weighted learning on the ADNI dataset to tailor classification models for personalized and cost-effective diagnosis. Their approach demonstrated a reduction in required biomarkers without compromising accuracy, illustrating the feasibility of AI in early detection.

In epidemiological studies, Witten et al. analyzed the impact of meteorological factors on Hand, Foot, and Mouth Disease (HFMD) in Wuwei City, China. Their use of regression analysis and weather data from 2008–2010 highlighted seasonal disease trends and informed public health interventions.

Another notable contribution is in plant pathology, where a spectral disease index was developed to detect and classify stages of wheat leaf rust. By leveraging reflectance spectra, this method enabled precise monitoring of disease severity levels, emphasizing the role of remote sensing and AI in agriculture.

In cardiac healthcare, several studies demonstrated the potential of machine learning for heart disease diagnosis. Quantized diagnosis methods using cardiac sound waveform analysis, as proposed by Jack et al., and non-linear heart rate variability measures were investigated for coronary heart disease (CHD) prediction.

III.OBJECTIVE

The analysis accuracy is reduced when the quality of medical data in incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. However, those existing work mostly considered structured data. There is no proper methods to handle semi structured and unstructured. The proposed system will consider both structured and unstructured data. The analysis accuracy is increased by using Machine Learning algorithm. In this work, our goal is to provide a tool to assist professionals and consumers in finding and choosing disease. To achieve this goal, we develop an approach that allows a user to query for disease that satisfy a set of conditions based on disease properties, such as disease indications and also takes into account patient profiles.

IV. PROPOSED METHODOLOGY

The proposed methodology aims to build a disease prediction system using machine learning techniques that can handle both structured and unstructured healthcare data. The steps involved are:
Data Collection: Healthcare datasets including symptoms, patient profiles, and diagnosis records are collected from reliable sources.

Data Preprocessing:
The data is cleaned by handling missing values and irrelevant entries. Structured data is normalized, and unstructured data (if any) is transformed into usable formats.

Feature Selection:
Relevant features are selected based on statistical methods and domain knowledge to reduce dimensionality and improve model accuracy.

Model Training:
A logistic regression algorithm is used to train the model. The data is split into training and testing sets, and the model is trained to predict diseases based on symptoms.

Disease Prediction:
Users input their symptoms into the system. The trained model processes these inputs and predicts the most probable diseases based on similarity to the training data.

Recommendation System:
Based on the predicted disease, the system provides risk levels, health suggestions, and next-step recommendations.

Model Evaluation:
The model's performance is evaluated using accuracy, precision, recall, and confusion matrix. The system achieved a high accuracy of 94.8%.

User Interface (Optional):
A simple interface is developed using Python Flask to allow users to interact with the model an review prediction results.

V. RESULT

The proposed system for disease prediction was successfully implemented using Python and machine learning algorithms, specifically Logistic Regression. The system was evaluated using real-world healthcare datasets, including information on COVID-19, Chronic Kidney Disease, and Heart Disease.
Key Results:
Accuracy:
The system achieved an overall prediction accuracy of 94.8%, outperforming several typical machine learning classifiers in both precision and convergence speed.
Functionality:
Users can input symptoms through a user interface. The system maps these symptoms to the most likely diseases by comparing them with the training dataset and returns probable diagnoses.

Prediction Capability:
The system is capable of predicting multiple diseases based on user symptoms and shows a percentage-based likelihood for each predicted disease.

Robust Data Handling:
The system processes both structured and unstructured medical data, increasing prediction robustness and making it more adaptable to real-world healthcare environments.

Visualization and Output:
Screenshots from the application demonstrate clear step-by-step data handling—from importing libraries and loading datasets to symptom-based prediction results. The system provides a user-friendly output displaying the disease prognosis alongside predicted conditions.

VI.CONCLUSION

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

REFERENCES

[1] A.Singh et al., "Heart Disease Prediction Using Machine Learning Algorithms", 2020 International Conference on Electrical and Electronics Engineering (ICE3), pp. 452-457, February 2020.

[2] A Narin, C Kaya and Z. Pamuk, "Automatic detection of coronavirus disease (COVID- 19) using x-ray images and deep convolutional neural networks", Mar 2020.

[3] Rajesh. Ranjan, "Predictions for COVID-19 outbreak in India using Epidemiological models", 2020. 14

[4] Mohan Senthilkumar, Chandrasegar Thirumalai and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", IEEE Access, vol. 7, pp. 81542-81554, 2019.

[5] Mohan Senthilkumar, Chandrasegar Thirumalai and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", IEEE Access, vol. 7, pp. 81542-81554, 2019.

[6] M. Bayati, S. Bhaskar and A. Montanari, "Statistical analysis of a low cost method for multiple disease prediction", Statistical Methods Med. Res., vol. 27, no. 8, pp. 2312-2328, 2018.

[7] A. K. Shrivas and S. Kumar Sahu, "Classification of Chronic Kidney Disease using Feature Selection Techniques", IJCSE, vol. 6, no. 5, pp. 649- 653, 2018.

[8] N Chaithra and B Madhu, "Classification Models on Cardiovascular Disease Prediction using Data Mining Techniques", Journal of Cardiovascular Diseases & Diagnosis, vol. 6, pp. 1-4, 2018.

[9] Kononenko I, Šimec E, Robnik-Šikonja M Artificial Intelligence in Medicine Machine learning for medical diagnosis: history, state of the art and perspective.