IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Disease Prediction Using Supervised Machine Learning

[1] Ketan Adling, [2] Aditya Date, [3] Nilesh Patil, [4] Saurabh Rathod, [5] Dr. Ashwini Shahapurkar

[1]Student, [2]Student, [3]Student, [4]Student, [5] Assistant Professor
[1]Department of Computer Engineering,
[1]G H Raisoni College of Engineering and Management, Pune, India

Abstract: This research explores the application of supervised machine learning to predict multiple diseases from patient-provided symptoms. The proposed system aims to assist in early disease detection, potentially reducing healthcare costs by enabling initial assessments without requiring immediate hospital visits. The system utilizes algorithms to analyze patient symptoms against a pre-existing dataset to predict the likelihood of diseases such as COVID-19, Chronic Kidney Disease, and Heart Disease.

Keywords: Machine Learning, Supervised Learning, Disease Prediction, Healthcare, Logistic Regression.

I. INTRODUCTION

In an era marked by increasing technological advancement, there is a growing demand for intelligent and accurate systems across various domains. While internet usage has become prevalent, individuals may inadvertently overlook their physical health. Often, minor health problems are ignored, which can escalate into serious diseases over time. Recognizing the potential of technology in addressing healthcare challenges, this research aims to develop a system capable of predicting multiple diseases based on symptoms reported by patients, thereby potentially reducing the necessity for immediate hospital or physician visits for initial diagnosis.

Machine Learning (ML), a subfield of Artificial Intelligence (AI), focuses on the development of algorithms that improve their performance through the analysis of data and experience. The process of machine learning typically involves two distinct phases: training and testing. Machine learning offers an efficient platform within the medical field to address diverse healthcare issues at an accelerated pace.

Within machine learning, two primary categories exist: Supervised Learning and Unsupervised Learning. In supervised learning, a model is constructed using well-labeled data, where each data point is associated with a known outcome. Conversely, unsupervised learning involves training models on unlabeled data, where the algorithm must identify patterns and structures without explicit guidance.

The core intent of this paper is to identify a suitable machine learning algorithm, specifically within the realm of supervised learning, that demonstrates both efficiency and accuracy in predicting diseases based on presented symptoms. The subsequent sections will detail the methodology and approach employed in this research.

II. SUPERVISED MACHINE LARNING FOR DISEASE PREDICTION

This data consists of input features (in this case, patient symptoms) and corresponding output labels (the diagnosed disease). The algorithm's objective is to learn a mapping function that can accurately predict the output label for new, unseen input data. The process involves training a model on a portion of the labeled data and then evaluating its performance on a separate, unseen portion of the data to assess its generalization ability. This research leverages the principles of supervised learning to build a predictive model for multiple diseases based on patient-provided symptoms. Supervised learning is a machine learning paradigm where an algorithm learns from labeled training data.

III. 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 5 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 6 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.

IV. OBJECTIVES

a) Aim of the Project:

- 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.

b) Scope of the Project

- 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.

V. METHODOLOGY

This research will involve the following key steps:

- Data Acquisition: Gathering a comprehensive dataset of patient symptoms and their corresponding diagnosed diseases. This dataset will serve as the labeled data required for supervised learning.
- Data Preprocessing: Cleaning and preparing the collected data to ensure its suitability for training a machine learning model. This may involve handling missing values, encoding categorical variables (symptoms and diseases), and scaling or normalizing numerical features if applicable.
- Feature Selection/Engineering: Identifying the most relevant symptoms (features) that contribute significantly to disease prediction. This step may involve using statistical techniques or domain expertise to select or create new features from the existing symptom data.
- Model Selection: Choosing an appropriate supervised machine learning algorithm for the disease prediction task. Based on the characteristics of the data and the research objectives, potential algorithms may include Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forests, or Naive Bayes.
- Model Training: Training the selected machine learning model using the preprocessed and feature-engineered dataset. The training process involves feeding the labeled data to the algorithm, allowing it to learn the relationships between symptoms and diseases.
- Model Evaluation: Assessing the performance of the trained model using a separate testing dataset that was not used during the training phase. Evaluation metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) will be used to quantify the model's predictive capabilities.
- System Development (Conceptual): Outlining the conceptual architecture of a system that would allow patients to input their symptoms, which would then be processed by the trained machine learning model to generate potential disease predictions.

VI. EXPECTED OUTCOMES

This research anticipates the development of a supervised machine learning model capable of predicting multiple diseases with a reasonable level of accuracy based on patient-reported symptoms. The findings will contribute to the understanding of how machine learning techniques can be effectively applied for early disease detection and preliminary health assessments. The conceptual system design may also provide a foundation for future development of user-friendly healthcare applications.

VII. Limitations

The It is important to acknowledge potential limitations of this research. The accuracy of the predictive model will heavily depend on the quality and comprehensiveness of the training data. Furthermore, self-reported symptoms may be subjective and potentially inaccurate. The model's predictions should not be considered a substitute for professional medical diagnosis and consultation.

VIII. Conclusion

This research endeavors to demonstrate the potential of supervised machine learning in developing a system for predicting multiple diseases based on patient symptoms. By leveraging labeled data and appropriate machine learning algorithms, the aim is to create a tool that can assist in early disease detection and empower individuals to seek timely medical attention when necessary. Further research and development will be crucial to validate and refine such systems for real-world healthcare applications.

IX. References

[1] 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] 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.

[3] 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.

[4] 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.