International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

A Machine Learning Method for Predicting Disease Based on Symptoms

Dr.G.Prabu (Associate Professor)#1 ,V.Prithvi Ramana#2, M.Mohamed Arsath #3, V.M.Sriram #4

COMPUTER SCIENCE AND ENGINEERING, SSM INSTITUTE OF ENGINEERING AND TECHNOLOGY
(Anna University)
1vgprabu.samy@gmail.com
2ramanaprithvi@gmail.com
3mdarsath2002@gmail.com
4infantsriram@gmail.com

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. A robust machine learning model that can efficiently predict the disease of a human, based on the symptoms that he/she possess. This research work carried out demonstrates the disease prediction system developed using Machine learning algorithms such as Decision Tree classifier. The paper presents the comparative study of the results of the above algorithms used. The disease is predicted using algorithms and comparison of the datasets with the symptoms provided by the user. The accurate analysis of medical database benefits in early disease prediction, patient care and community services. The techniques of machine learning have been successfully employed in assorted applications including Disease prediction. .

I. INTRODUCTION

MACHINE LEARNING A SUBFIELD OF ARTIFICIAL INTELLIGENCE THAT INVOLVES TRAINING COMPUTERS TO LEARN IS FROM DATA, WITHOUT BEING EXPLICITLY PROGRAMMED. MACHINE LEARNING CAN BE APPLIED TO A WIDE RANGE OF PROBLEMS, INCLUDING DISEASE PREDICTION. DISEASE PREDICTION INVOLVES USING AVAILABLE DATA ABOUT A PATIENT, SUCH AS THEIR MEDICAL SYMPTOMS, AND DEMOGRAPHICS, TO PREDICT THE LIKELIHOOD THAT THEY HAVE A PARTICULAR DISEASE. MACHINE LEARNING ALGORITHMS CAN BE USED TO ANALYZE THIS DATA AND GENERATE A PREDICTIVE MODEL THAT CAN BE USED TO MAKEDISEASE PREDICTION.THERE ARE SEVERAL STEPS INVOLVED IN DEVELOPING A MACHINE LEARNING MODEL FOR DISEASE PREDICTION. THE FIRST STEP IS TO COLLECT AND CLEAN THE DATA. THIS INVOLVES IDENTIFYING RELEVANT DATA SOURCES AND ENSURING THAT THE DATA IS ACCURATE AND COMPLETE. ONCE THE DATA HAS BEEN COLLECTED, IT IS TYPICALLY PREPROCESSED TO TRANSFORM IT INTO A FORMAT THAT CAN BE USED BY MACHINE LEARNING ALGORITHMS.

A. *EXISTING SYSTEM*

The existing system is Lyme disease can be prevented if antibiotic prophylaxis is given to a patient within 72 hours of a blacklegged tick bite.Therefore, recognizing a blacklegged tick could facilitate the management of Lyme disease. Methods: In this work, we build an automated detection tool that can differentiate blacklegged ticks from other tick species using advanced computer vision approaches in real-time.Convolution neural network model achieves 92% accuracy on unseentick species. Conclusion: Our proposed vision-based approach simplifies tick identification.Geography of exposure and potentially be leveraged to inform the risk of Lyme disease infection. This is the first report of using deep learning technologies to classify ticks, providing the basis for automation of tick surveillance**,**

1. DRAWBACKS

- **Data Quality and Availability:** The effectiveness of symptom-driven disease prediction heavily relies on the quality and availability of data. Incomplete or inaccurate symptom reportingng can lead to biased predictions or false alarms.

- **Limited Scope:** Symptom-driven approaches may not capture all relevant factors influencing disease outcomes, such as genetic predispositions, environmental factors, or lifestyle choices.This limitation can affect the accuracy and comprehensiveness of predictions.

- **Ethical and Privacy Concerns:** The use of personal health data for predictive analytics raises ethical concerns regarding patient privacy, data security, and consent. Ensuring compliance with regulations like HIPAA (Health Insurance Portability and Accountability Act) is essential but challenging.

B. ***PROPOSED SYSTEM***

Machine learning can be used to predict diseases using decision tree algorithms.

Decision tree algorithms are a type of supervised learning algorithm that is used for classification and regression tasks.

They build a model in the form of a tree structure, where each node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label or a numeric value.

To predict diseases using decision tree algorithms, you will need a dataset that contains features related to the disease, such as symptoms. Then predict the disease.

1.ADVANTAGES

Interpretability: Decision trees are easy to understand and interpret, making them useful for healthcare professionals and patients to comprehend how a prediction is made based on the presence or absence of certain symptoms or features.

Handling Non-linear Relationships: Decision trees can capture non-linear relationships between features and the target variable, allowing them to effectively model complex disease patterns that may not be easily discernible using linear methods.

Feature Selection: Decision trees inherently perform feature selection by determining which features are most informative for making predictions. This can help in identifying the most relevant symptoms or risk factors associated with a disease.

Handling Missing Values: Decision trees can handle missing values in the dataset without requiring imputation techniques. They simply consider the available information at each node to make decisions, making them robust to incomplete data.

Scalability: Decision tree algorithms can handle large datasets efficiently, making them suitable for analyzing extensive healthcare databases containing information from numerous patients. Additionally, they can be easily parallelized to further enhance scalability

II. LITERATURE SURVEY

1.Geumkyung Nah Eun-A Choi et all proposed "Gene expression analysis known COPD loci revealed its varied levels by disease severity"IEEE-2021

Numerous genome studies on chronic obstructive pulmonary disease (COPD) have revealed hundreds of associated loci. However, understanding of COPD pathogenesis in the context of biological pathway, especially for disease severity, has remained unclear. In this study, we studied gene expression patterns of known COPD loci among mild, moderate, and severe COPD patients. Although expression levels of 16 known genes available from RNA-seq were not significantly different across COPD severity by comparing varied levels of COPD severity ($P>0.05$), DIP2B showed higher gene expression mean changes in mild than severe group. Taken together, this study demonstrated different gene expression changes by COPD severity among known COPD loci. The results will be valuable scientific evidence to enhance our understanding of COPD risk, severity, and related pathogenesis.

2. **Hiroki Fuse Kota Oishi et all proposed "Detection of Alzheimer's Disease with Shape Analysis of MRI Images"IEEE-2019**

In the current study, we tested the effectiveness of a method using brain shape information for classification of healthy subjects and Alzheimer's disease patients. A P-type Fourier descriptor was used as shape information, and the lateral ventricle excluding the septum lucidum was analyzed. Using a combination of several descriptors as features, we performed classification using a support vector machine. The results revealed classification accuracy of 87.5%, which was superior to the accuracy achieved using volume ratio to intracranial volume (81.5%), which is widely used for conventional evaluation of morphological changes. The current findings suggest that shape information may be more useful in diagnosis, compared with conventional volume ratio.

III. REQUIREMENT AND ANALYSIS

HARDWARE REQUIREMENTS

- Processor : Minimum Intel i5
- RAM : Min 8 GB
- Hard Disk : 500 GB

SOFTWARE REQUIREMENTS

- *Operating System : WINDOWS 10/11*
- *Language Used :PYTHON*
- *Back End : PYTHON IDEL*
- *Front End :PYTHON SHELL*

SYSTEM ARCHITECTURE

Data Acquisition:

Data from various sources such as electronic health records (EHRs), medical databases, patient-reported symptoms, wearable devices, and health apps need to be collected.APIs or data pipelines can be used to fetch real-time data from sources like hospitals, clinics, or IoT devices.

Data Preprocessing:

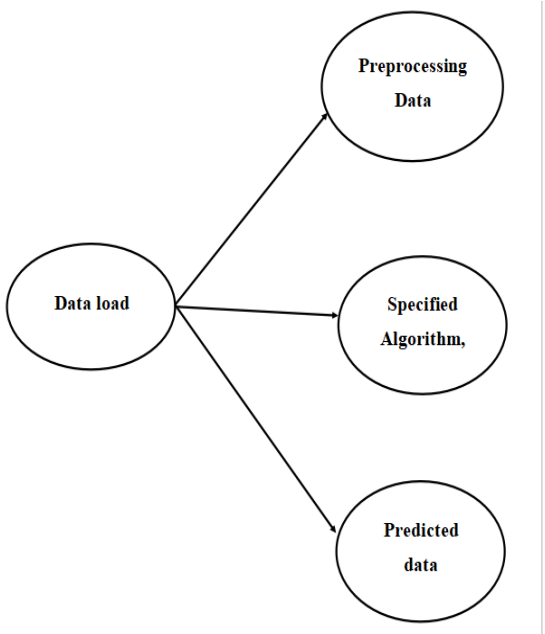
Clean and preprocess the collected data to handle missing values, outliers, and inconsistencies.Perform data normalization, encoding categorical variables, and feature scaling as necessary. Handle privacy and security concerns related to medical data, ensuring compliance with regulations such as HIPAA.

Feature Extraction:

Extract relevant features from the preprocessed data that can be used to represent symptoms and other health-related information. Techniques such as dimensionality reduction (e.g., PCA) and feature selection (e.g., mutual information, feature importance) can be applied to reduce the feature space and improve model performance

USECASE DIAGRAM

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a systemand the different use cases and will often be accompanied by other types of diagrams as well. The use cases are represented by either circles or ellipses.



IV.DESIGN & MODULE DESCRIPTION

Systems implementation is the process of: defining how the information system should be built (i.e., physical system design), ensuring that the information systemic operational and used, ensuring that the information system meets quality standard (i.e., quality assurance).Implementation is the process that actually yields the lowest-level system elements in the system hierarchy (system breakdown structure). System elements are made, bought, or reused. Production involves the hardware fabrication processes of forming, removing, joining, and finishing, the software realization processes of coding and testing, or the operational procedures development processes for operators' roles. If implementation involves a production process, a manufacturing system which uses the established technical and management processes may be required. The purpose of the implementation process is to design and create (or fabricate) a system element conforming to that element's design properties and/or requirements.

A. ***MODULES***

1)DATASET COLLECTION
2)DATA PREPROCESSING
3)FEATURE EXTRACTION
4)PREDICTED THE DATA
5)CLASSIFY DATA

1) *MODULE DESCRIPTION*

1. DATASET COLLECTION:

A machine learning datasets is a collection of data in fake reviews that is used to train the model. A datasets acts as an example to teach the machine learning algorithm how to make predictions. Datasets contains a lot of separate pieces of data but can be used to train an algorithm with the goal of finding predictable patterns inside the whole datasets in machine learning.

In this section, you create a Dataset using the donation file. When numerical data is loaded, it's treated as a measure. You, also, learn how to correct the Treat as value for numerical columns that are attributes.

2. **DATA PREPROCESSING:**

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and

crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task.

3. FEATURE EXTRACTION:

Feature extraction refers to the process of transforming CSV data into numerical and string features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the data. Feature extraction can be accomplished manually or automatically: Manual feature extraction requires identifying and describing the features that are relevant for a given problem and implementing a way to extract those features.

4. PREDICTED THE DATA:

Predicting data using machine learning involves a combination of data preparation, model selection, training, evaluation, and prediction. It is an iterative process that may involve multiple rounds of experimentation and fake review refinement to the command in processing data and predicted the output graph also predicting dataset.

5. CLASSIFY DATA

In machine learning prediction, the dataset is typically split into a training set and a testing set. The training set is used to train the machine learning model, while the testing set is used to evaluate the performance of the model. Some datasets, the number of examples for each class or category may be balanced, while in others, there may be an

imbalance in the number of examples. Imbalanced data can pose a challenge for some machine learning algorithms, as they may struggle to learn from the minority class.

V. IMPLEMENTATION

A.*FRONT END*

1.PYTHON SHELL :

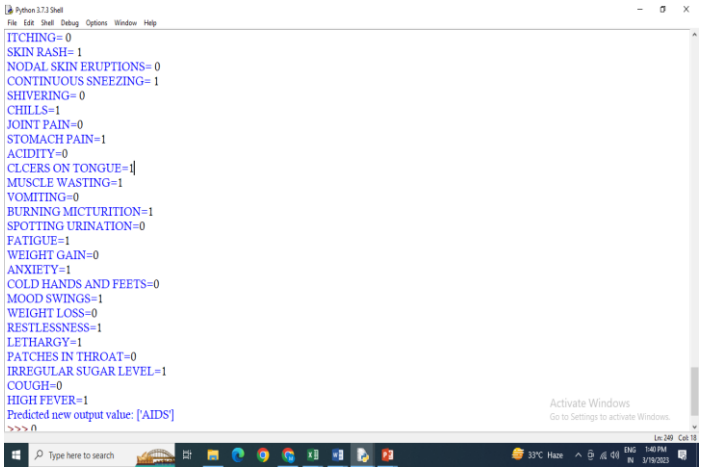
The Python shell, also known as the Python interactive interpreter or Python REPL (Read-Eval-Print Loop), offers an intuitive and interactive environment for executing Python code line by line. Upon launching the shell, typically initiated by entering 'python' or 'python3' in the terminal or command prompt, users are greeted with a prompt (usually '>>>') where Python code can be entered directly. This interactive mode allows for real-time experimentation with Python syntax, enabling users to test small code snippets, explore Python libraries, and evaluate expressions dynamically. Operating within a loop known as the Read-Eval-Print Loop (REPL), the Python shell reads user input, evaluates the provided expressions or statements, prints the result (if any), and then loops back to await further input. Users can access help and documentation for Python functions, modules, and objects directly within the shell using the 'help()' function or by appending '?' followed by the object of interest. Furthermore, the shell supports multiline input, where users can continue entering code until a complete statement is formed. Exiting the Python shell is straightforward, accomplished by typing 'exit()' or pressing Ctrl + D (Ctrl + Z on Windows), allowing users to seamlessly transition back to the command prompt or terminal. Overall, the Python shell serves as a valuable tool for quick code prototyping, debugging, and interactive learning, offering a convenient space for exploring Python's capabilities and features.

B. BACK END

1.PTYTHON IDLE:

Python IDLE, an abbreviation for "Integrated Development and Learning Environment," is an integrated development environment (IDE) for Python programming language beginners and professionals alike. It offers a user-friendly interface with features designed to facilitate coding, testing, and debugging Python scripts. Upon launching IDLE, users are greeted with a customizable editor window where they can write Python code. The editor provides syntax highlighting, auto-indentation, and code completion features, enhancing code readability and productivity. Additionally, IDLE includes an interactive Python shell, enabling users to execute Python code interactively and experiment with language features in real-time. The shell also serves as a valuable tool for debugging code snippets and exploring Python libraries interactively. IDLE further supports the creation and management of Python scripts, allowing users to save, open, and run Python files directly from the IDE. Its simplicity and ease of use make it an ideal choice for beginners learning Python programming, while its robust features cater to the needs of more experienced developers, making it a versatile and widely used development environment in the Python community.

A.SCREENSHOTS



VI CONCLUSION AND FUTURE WORK

Disease prediction through machine learning, especially leveraging decision tree algorithms, represents a pivotal advancement in healthcare diagnostics. By synthesizing and analyzing vast amounts of medical data, these algorithms empower healthcare professionals with valuable insights, potentially leading to more accurate and timely diagnoses. However, it's crucial to acknowledge that while machine learning models offer significant advancements, they are not infallible and should complement rather than replace clinical judgment and expertise. The ongoing evolution of this field promises continued enhancements in disease prediction methodologies, with interdisciplinary collaborations and technological advancements driving innovation forward. Through continuous research and development efforts, we anticipate further refinements in predictive accuracy, model interpretability, and scalability, ultimately contributing to enhanced healthcare outcomes and improved patient care globally. As we navigate this ever-evolving landscape, it's imperative to embrace a collaborative approach, where machine learning augments human expertise, fostering a symbiotic relationship between technology and healthcare professionals for the betterment of public health

VII. REFERENCES

[1] K. Kugeler, A. Schwartz, M. Delorey, P. Mead, and A. Hinckley, ‘‘Estimating the frequency of Lyme disease diagnoses, United States, 2010–2018,’’ Emerg. Infectious Disease J., vol. 27, no. 2, p. 616, 2021.

[2] A. Schwartz, K. Kugeler, C. Nelson, G. Marx, and A. Hinckley, ‘‘Use of commercial claims data for

evaluating trends in Lyme disease diagnoses, United States, 2010–2018,’’ Emerg. Infectious Disease J., vol. 27, no. 2, p. 499, 2021.

[3] Public Health Agency of Canada. (2021). Surveillance of Lyme Disease. Accessed: May 5, 2021. [Online]. Available: https://www.canada.ca/en/ public-health/services/diseases/lyme-disease/surveillance-lymedisease.html#a3

[4] Z. S. Y. Wong, J. Zhou, and Q. Zhang, ‘‘Artificial intelligence for infectious disease big data analytics,’’ Infection, Disease Health, vol. 24, no. 1, pp. 44–48, Feb. 2019.

[5] A. C. Miller, I. Singh, E. Koehler, and P. M. Polgreen, ‘‘A smartphonedriven thermometer application for real-time population- and individuallevel influenza surveillance,’’ Clin. Infectious Diseases, vol. 67, no. 3, pp. 388–397, Jul. 2018.

[6] D. Hendrycks, K. Lee, and M. Mazeika, ‘‘Using pre-training can improve model robustness and uncertainty,’’ 2019, arXiv:1901.09960.

[7] C. Lam, D. Yi, M. Guo, and T. Lindsey, ‘‘Automated detection of diabetic retinopathy using deep learning,’’ AMIA Summits Transl. Sci., vol. 2018, no. 1, p. 147, 2018.

[8] S. Akbarian, N. M. Ghahjaverestan, A. Yadollahi, and B. Taati, ‘‘Noncontact sleep monitoring with infrared video data to estimate sleep apnea severity and distinguish between positional and nonpositional sleep apnea: Model development and experimental

validation,'' J. Med. Internet Res., vol. 23, no. 11, Nov. 2021, Art. no. e26524.

[9] S. Akbarian, L. Seyyed-Kalantari, F. Khalvati, and E. Dolatabadi, ''Evaluating knowledge transfer in neural network for medical images,'' 2020, arXiv:2008.13574.

[10] L. Wang and K.-J. Yoon, ''Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks,'' 2020, arXiv:2004.05937.