A Multi-Model Machine Learning Framework for Accurate Disease Prediction

K.Pavani[1], A.Sivalatha[2]
#1 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

#2 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

Abstract: Machine learning can diagnose and treat many diseases. Powerful multiple machine learning algorithms provide predictive analysis to improve patient care and illness prognosis. With machine learning, disease prediction may be done swiftly and accurately. Clinical criteria and machine learning categorize diseases. We conclude with a graph comparing machine learning-based classification methods. Machine learning may assist predict all diseases in addition to medical therapy. Using predictive analysis and efficient multiple machine learning algorithms helps provide more accurate illness forecasts, improving patient care. Fast and accurate sickness prediction is possible with machine learning techniques. Clinical characteristics are used to classify diseases using machine learning algorithms. Finally, we graph the outcomes of several machine learning classification methods

Index Terms *- Machine Learning, Disease Prediction, Healthcare Analytics, Symptom-Based Diagnosis, Naive Bayes, Decision Tree, Support Vector Machine (SVM), , Data Mining*

1. INTRODUCTION

Since each illness has its own symptoms, patient health demands ongoing attention. In conventional medicine, symptoms are prioritized to speed up diagnosis and accurately predict illness existence. The illness-affecting symptom is prioritized. Data mining aids disease diagnosis and prognosis by supplementing and replacing medical expertise. This procedure allows us to extract information, correlations, and patterns that standard statistical methods cannot. Always having the appropriate conclusion and privilege helps therapy work. Data mining principles can reveal hidden statistics, database links, and machine learning approaches to further analyze a patient's status based on clinical information. Machine learning algorithms forecast many diseases and provide effective treatment plans to promote health while lowering the risks of high costs, sluggish recovery, and improper therapy. Fortunately, machine learning algorithms are effective at illness prediction, although many methods remain unexplored. This study presented a graphical user interface (GUI) application that employs machine learning to forecast diseases by mining symptoms and comparing methods' efficacy and accuracy. This software detects diseases. We will achieve our aim using Naive Bayes, Decision Trees, Support Vector Machines, and MLP. New insertions' results can be accurately anticipated by these methods. Examine the data, both graphical and numerical, that reveals the universal trend throughout all training programmes. We also use a graphical user interface programme to determine if a patient will be afflicted with a disease. The goal of the proposed framework is to solve the problem of accurate disease prediction for a given patient by immediately collecting appropriate data and components associated with various diseases and separating out crucial features that could help us verify the typical examples when selecting the model and its parameters. Briefly, the presentation is broken down as follows: first, we review fundamental machine learning tactics for disease prediction; second, we analyse the research design, data, and methods employed. Then, we get into the conclusions we draw at the very end.

2. LITERATURE SURVEY

2.1 Issues in sustainable transportation.

Sustainability, sustainable development, and sustainable transportation are gaining popularity. This essay highlights problems with the definition, assessment, and application of sustainable transportation. The variety of definitions of sustainability, the variety of issues that fall under the umbrella of sustainability, the variety of viewpoints, the criticism of sustainability analysis, the evaluation of sustainability, the impact of transportation on sustainability, goals versus objectives, sustainable transportation decision-making, automobile dependency, equity, land use, community liveability, and sustainable transportation solutions are all significant issues. Originally concentrating on a few concerns related to resource use, sustainable development is now more widely defined to encompass human health, equality, economic and social wellbeing, and ecological integrity. A more comprehensive definition of sustainable transport tends to favor more integrated solutions, such as better travel options, financial incentives, institutional reforms, land use changes, and technical innovation, whereas a more restrictive definition tends to favor specific technological solutions. Planning for sustainability may necessitate altering how individuals approach and resolve transportation-related issues.

2.2 A technical solution to improve the red cab for touring in Chiang Mai: Chinese tourists' perspective

Thailand's most popular tourist destination, particularly for Chinese travelers, is Chiang Mai. An essential part of the tourist sector is transportation [1]. However, the most annoying aspect of traveling in Chiang Mai for foreigners is having to use local public transportation, which has no set routes and constantly necessitates direct contact between visitors and the driver. As a result, issues that might hinder the city's tourism growth may arise. For instance, visitors may easily get lost if they are dropped off at the incorrect location as a result of misunderstandings. In order to improve visitor satisfaction from a technology standpoint, this article focuses on suggesting a technical solution to improve the present local transportation system. According to an assessment, the solution is thought to significantly enhance Chiang Mai's public transportation system.

2.3 GPS based Public Transport Arrival Time Prediction:

lengthy waits at bus stations and frequently no bus arrival after a lengthy wait are only two of the many issues that public transportation users in developing nations deal with. Displaying the anticipated bus arrival times at the appropriate bus stops is one of the creative solutions to the issue. Numerous prediction systems have been created using various methods.Instead of using antiquated methods, this study proposes a public transportation arrival time prediction system that uses real-time Automatic Vehicle Location (AVL) data. Real-Time Locating Systems (RTLS) and communication across various nodes are developed using the Global Positioning System (GPS) and Global System for Mobile Communications (GSM), respectively. An effective and optimized method is created on the server side to precisely forecast the arrival timings. On a chosen route, many test drives were conducted under varied test conditions. Additionally, graphs at the end of the study display the resulting anticipated timings and their variation from Final Traveling Time (FTT) for each bus stop.

2.4 Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015:

One of the main causes of the world's illness burden, exposure to ambient air pollution raises morbidity and death. From 1990 to 2015, we looked at global, regional, and national changes in mortality and illness burden related to ambient air pollution. METHODS: Using satellite-based estimations, chemical transport models, and ground-level data, we calculated the worldwide population-weighted mean concentrations of ozone and particle mass with aerodynamic diameter less than 2.5 mum (PM2.5) at a resolution of around 11 km x 11 km. We used non-linear exposure-response functions covering the global range of exposure to estimate the relative risk of mortality from ischemic heart disease, cerebrovascular disease, chronic obstructive pulmonary disease, lung cancer, and lower respiratory infections from epidemiological studies using integrated exposure-response functions for each cause of death. FINDINGS: In 2015, ambient PM2.5 ranked as the seventh most important death risk factor. In 2015, exposure to PM2.5 resulted in 103.1 million (90.8 million 115.1 million) disability-adjusted life-years (DALYs) and 4.2 million (95% uncertainty interval [UI] 3.7 million to 4.8 million) deaths worldwide, accounting for 7.6% of all deaths and 4.2% of all DALYs, with 59% of these occurring in east and south Asia. Between 1990 and 2015, the number of deaths linked to ambient PM2.5 rose from 3.5 million (95% UI 3.0 million to 4.0 million) to 4.2 million (3.7 million to 4.8 million). In 2015, ozone exposure resulted in an extra 254 000 (95% UI 97 000-422 000) deaths and 4.1 million (1.6 million to 6.8 million) DALYs from chronic obstructive pulmonary disease. INTERPRETATION: Due to population aging, shifting rates of non-communicable diseases, and rising air pollution in low-income and middle-income nations, ambient air pollution significantly contributed to the global burden of illness in 2015, which has risen during the previous 25 years. The most polluted nations will only see modest burden reductions until PM2.5 levels are significantly lowered, although exposure reduction may have a significant positive impact on health. FUNDING: Health Effects Institute and the Bill & Melinda Gates Foundation.

2.5 Design of Taxi Vehicle Monitoring System Based on GPS/GPRS

A type of vehicle monitoring terminal device that combines GPS, GPRS, and LCD and uses the C8051F040 microcontroller as its core was suggested in response to the requirements of the security operation taxi. This gadget will demonstrate the supremacy of the C8051F040 microcontroller, which has a comprehensive interface, low power consumption, and fast speed. Vehicle localization, vehicle monitoring and dispatching, and other tasks may be realized by the device through the use of GPS, LCD, and the GPRS module that communicates with the monitor center. This document provides a detailed description of the hardware design and accompanying procedure flow process.

3. METHODOLOGY

i) Proposed Work:

As part of the proposed system, four machine learning algorithms are utilized to predict diseases using the developed graphical user interface (GUI) model. The implemented algorithms include Support Vector Machines (SVM), Decision Tree, Naive Bayes (Naïve Bayes), and Multi-Layer Perceptron (MLP). These machine learning classification techniques are compared based on their prediction accuracy and overall performance in disease detection.The proposed process consists of four major stages. The first stage involves selecting important clinical features required for disease prediction. In the second stage, all relevant symptoms and medical parameters are collected and organized for analysis. The third stage applies machine learning algorithms to classify diseases based on the extracted clinical characteristics and symptom information. Finally, the performance results of the implemented machine learning algorithms are represented graphically for comparative analysis and evaluation.The growing availability of electronic healthcare data has significantly improved the effectiveness of machine learning techniques in disease analysis and prediction. Modern computational methods have become increasingly advanced and capable of handling complex medical datasets efficiently, enabling more accurate and reliable disease prediction systems.

ii) System Architecture:

The proposed system architecture begins with a dataset containing patient details and symptoms, which serves as the primary input source. Initially, attribute selection is performed to identify the most relevant features that significantly influence disease prediction. This is followed by data preprocessing, where missing values, noise, and inconsistencies are handled to ensure clean and structured data. The processed data is then passed to the classification stage, where multiple machine learning algorithms such as Naive Bayes (NB), Support Vector Machine (SVM),MLP, and Decision Tree (DT) are applied to learn patterns from the data and build predictive models.

Once the models are trained, they are used to perform disease prediction based on new input symptoms. The predicted results from different algorithms are then evaluated using performance measures such as accuracy, precision, recall, and F1-score to determine the best-performing model. Finally, the results are visualized through graphs and user-friendly outputs, enabling easy interpretation of predictions. This architecture ensures an efficient pipeline from data input to result visualization, improving the accuracy and usability of the disease prediction system.

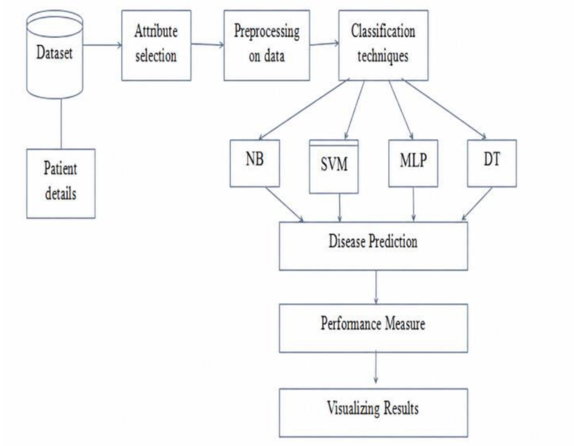

Fig 1 System Architecture

iii) MODULES:

1. Data Collection Module

- Collects dataset containing patient symptoms and disease labels
- Includes patient details from medical records or datasets

2. Attribute Selection Module

- Selects important features (symptoms) from dataset
- Removes irrelevant and redundant data to improve accuracy

3. Data Preprocessing Module

- Cleans data by handling missing values and noise
- Converts data into suitable format for machine learning

4. Model Training Module

- Trains multiple algorithms (Naive Bayes, SVM, MLP, Decision Tree)
- Splits dataset into training and testing sets

5. Classification Module

- Applies trained models to classify diseases
- Compares outputs from different algorithms

6. Prediction Module

- Predicts disease based on user input symptoms
- Provides real-time results

7. Performance Evaluation Module

- Calculates accuracy, precision, recall, and F1-score
- Identifies best-performing model

8. Visualization Module

- Displays results using graphs and charts
- Shows comparison of algorithm performance

v) ALGORITHMS:

1. Naive Bayes Algorithm

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, which assumes that all input features (symptoms) are independent of each other. It calculates the probability of a disease occurring given a set of symptoms and selects the class with the highest probability. Due to its simplicity and efficiency, it performs well on large medical datasets and is widely used for disease prediction where categorical data is involved.

2. Decision Tree Algorithm

Decision Tree is a supervised learning algorithm that represents data in the form of a tree structure, where each node corresponds to a feature (symptom), each branch represents a decision rule, and each leaf node represents the predicted disease. It splits the dataset based on conditions that maximize information gain, making it easy to understand and interpret. This algorithm is particularly useful in healthcare applications where decision transparency is important.

3. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful classification technique that identifies an optimal hyperplane to separate different classes in a high-dimensional feature space. It focuses on maximizing the margin between data points of different classes, which improves generalization and accuracy. SVM is effective for complex datasets and works well when the relationship between symptoms and diseases is non-linear.

4. Multilayer Perceptron (MLP) Algorithm Working

Multilayer Perceptrons (MLPs) are used in machine learning and deep learning for categorization and prediction. It is a popular neural network method because it can comprehend complicated data patterns and correlations. MLP is inspired by the human brain's connected neurons processing information and making choices. Many healthcare systems employ MLP to diagnose diabetes, cancer, illness, medical image analysis, and heart disease.

Three layers make up an MLP network. These levels are input, disguised, and output. This layer receives dataset data including patient age, blood pressure, cholesterol, glucose, and symptoms. One property or feature per input layer neuron. Hidden layers compute and locate data links. An MLP may have hidden layers depending on how difficult the job is. The output layer provides the final forecast, which may include illness detection.

Correct MLP algorithm functioning begins with forward propagation. This technique moves entered data from the input layer to the concealed layers before sending it to the output layer. Every neuron multiplies input values by their weights. A bias value is added to the weighted total. Procedure's.

4. EXPERIMENTAL RESULTS

The proposed disease prediction system was evaluated using a medical dataset containing various symptoms and corresponding diseases. The dataset was divided into training and testing sets to build and validate the machine learning models. Multiple classification algorithms, including Naive Bayes, Decision Tree, Support Vector Machine (SVM), and MLP, were applied and compared based on performance metrics such as accuracy, precision, recall, and F1-score. The evaluation results indicate that ensemble methods, particularly Decision Tree achieved higher accuracy and better generalization compared to other individual models.

The system also provides a graphical comparison of all algorithms, allowing easy visualization of performance differences. As shown in the implementation screenshots (Pages 45–48), the GUI enables users to input symptoms and obtain real-time predictions along with performance outputs. The experimental results demonstrate that the proposed system is efficient, reliable, and capable of providing accurate disease predictions, making it suitable for practical healthcare applications.

Accuracy: A test's accuracy is defined as its ability to recognize debilitated and solid examples precisely. To quantify a test's exactness, we should register the negligible part of genuine positive and genuine adverse outcomes in completely examined cases. This might be communicated numerically as:

Accuracy = TP + TN TP + TN + FP + FN.

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Precision measures the proportion of properly categorized occurrences or samples among the positives. As a result, the accuracy may be calculated using the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$Accuracy = \frac{(TN + TP)}{T}$$

Recall: Recall is a machine learning metric that surveys a model's capacity to recognize all pertinent examples of a particular class. It is the proportion of appropriately anticipated positive perceptions to add up to real up-sides, which gives data about a model's capacity to catch instances of a specific class.

$$Accuracy = \frac{(TN + TP)}{T}$$

F1-Score: The F1 score is a machine learning evaluation measurement that evaluates the precision of a model. It consolidates a model's precision and review scores. The precision measurement computes how often a model anticipated accurately over the full dataset.

$$Accuracy = \frac{(TN + TP)}{T}$$

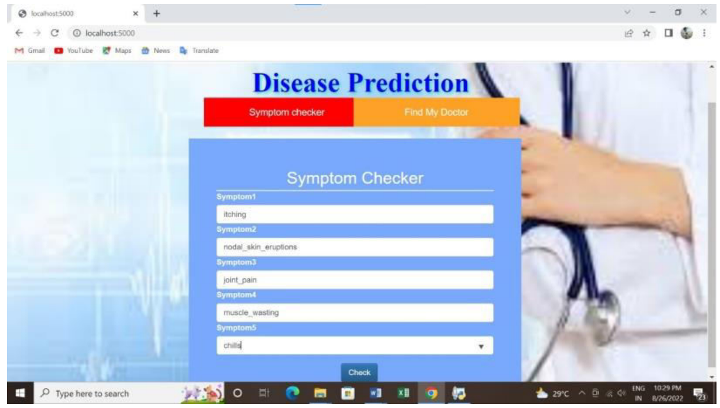

Fig 2:Input Data

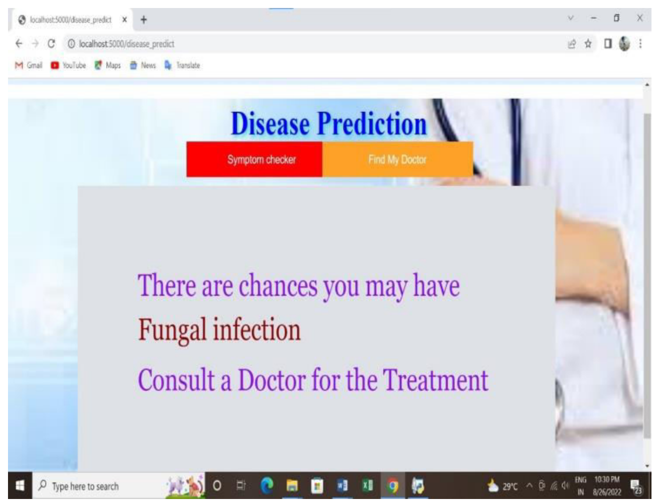

Fig 3:Predict Output

5. CONCLUSION

The proposed system demonstrates an effective machine learning-based approach for disease prediction using patient symptoms. By integrating multiple classification algorithms such as Naive Bayes, Decision Tree, Support Vector Machine, and MLP, the system improves prediction accuracy and reliability. The use of data preprocessing and feature selection further enhances model performance and ensures meaningful results.

The developed GUI makes the system user-friendly and enables real-time prediction, making it suitable for practical healthcare applications. Overall, the system reduces diagnostic time, minimizes human error, and supports medical professionals in decision-making. The experimental results confirm that the approach is efficient, scalable, and capable of improving early disease detection.

6. FUTURE SCOPE

The proposed disease prediction system can be further enhanced by integrating advanced deep learning techniques such as Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) to improve prediction accuracy for complex medical datasets. The system can also be extended by incorporating larger, real-time healthcare datasets from hospitals and medical institutions, enabling more reliable and scalable predictions.

Additionally, the application can be developed into a mobile or web-based platform for wider accessibility, allowing patients and doctors to use the system remotely. Integration with IoT-based health monitoring devices can enable continuous data collection and real-time disease prediction. Future improvements may also include personalized treatment recommendations, multilingual support, and enhanced security features for protecting sensitive patient data.

REFERENCES

[1] Thailand Motor Vehicle Registered. CEIC Data Global Database.

[2] Linda, S., Can public transport compete with the private car? Iatss Research, 2003. 27(2): p. 27-35.

[3] CO2 emissions (metric tons per capita) - Thailand. THE WORLD BANK.

[4] Cohen, A.J., et al., Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. The Lancet, 2017. 389(10082): p. 1907-1918.

[5] Litman, T. and D. Burwell, Issues in sustainable transportation. International Journal of Global Environmental Issues, 2006. 6(4): p. 331-347.

[6] Liu, M. and N. Choosri.. A technical solution to improve the red cab for touring in Chiang Mai: Chinese tourists' perspective. in 2016 Chinese Control and Decision Conference (CCDC). 2016. IEEE

[7] Farooq, M.U., A. Shakoor, and A.B. Siddique. GPS based Public Transport Arrival Time Prediction. in 2017 International Conference on Frontiers of Information Technology (FIT). 2017. IEEE.

[8] Bin, Y., Y. Zhongzhen, and Y. Baozhen, Bus arrival time prediction using support vector machines. Journal of Intelligent Transportation Systems, 2006. 10(4): p. 151-158.

[9] Maiti, S., et al. Historical data based real time prediction of vehicle arrival time. in 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). 2014. IEEE.

[10] Fan, W. and Z. Gurmu, Dynamic travel time prediction models for buses using only GPS data. International Journal of Transportation Science and Technology, 2015. 4(4): p. 353-366.

Author Profiles

Mrs.K.Pavani Working as Assistant & Head of Department of MCA ,in SRK Institute of technology in Vijayawada. She done with MCA and M. Tech in Computer Science .She has 10 years of Teaching experience in SRK Institute of technology, Enikepadu, Vijayawada,NTR District. Her area of interest includes AI ML, etc.

Ms.A.Sivalatha is an MCA Student in the Department of Computer Application at SRK Institute Of Technology, Enikepadu, Vijayawada, NTR District. He has Completed Degree in B.Sc.(Mathematics,Physics,Computer Science) from Raos Degree College Nandyala. His area of interest are DBMS and Machine Learning with Python.