International Journal of Advanced Engineering and Nano Technology (IJAENT)
ISSN: 2347-6389 (Online), Volume-10 Issue-11, November 2023

Early Disease Prediction using Ml

Amit Kumar, Harshika Bansal, Ayush Jaiswal, Sovit Kumar Gupta

Abstract: ***The approach employed in disease prediction using machine learning involves making forecasts about various diseases by utilizing symptoms provided by patients or other individuals. The supervised machine learning approaches called random forest classifier, KNN classifier, SVMs classifier are employed to forecast the disease. These algorithms are used to determine the disease's probability. Accurate medical data analysis helps with patient care and early disease identification as biomedical and healthcare data volumes rise. Diabetes, heart diseases are just a few of the illnesses we can forecast using linear regression and decision trees. Early detection is beneficial for determining the possibility of diabetes, heart disease.***

Keywords: Machine Learning, classifiers, probability, prediction, approaches etc..

I. INTRODUCTION

1.1 Introduction

Machine learning has found extensive applications across diverse domains, the advancement of technology, along with improved computational power and the availability of open-source datasets, has had a significant impact on the extensive adoption of machine learning (ML) in various fields, such as education and healthcare. Within the healthcare industry, ML plays a pivotal role in analyzing extensive datasets encompassing medical images and patient records to detect patterns and offer predictions. Its utilization has proven invaluable in tackling a range of healthcare-related obstacles [6].

For instance, heart disease manifests differently in individuals, with varying degrees of severity. By developing a machine learning model and training it on relevant datasets, healthcare professionals can input individual patient details to predict the progression of the disease[2][11][12][13][14]. These predictions are tailored to the specific characteristics of each patient [6]. Another health concern, type-2 diabetes, can be prevented through measures like weight control and lifestyle adjustments. In this particular study, machine learning models are employed to predict the probability of various illnesses, including but not limited to the coronavirus, heart disease, and diabetes [4]. The risk of the diseases is predicted within seconds.

Manuscript received on 09 July 2023 | Revised Manuscript received on 09 November 2023 | Manuscript Accepted on 15 November 2023 | Manuscript published on 30 November 2023.
*Correspondence Author(s)
Prof. Amit Kumar, Professor, Galgotias University Greater Noida, UP, India. amit.kr@galgotiasuniversity.edu.in
Harshika Bansal, Student, Galgotias University Greater Noida, UP, India. harshika.20scse1010558@galgotiasuniversity.edu.in
Ayush Jaiswal*, Student, Galgotias University Greater Noida, UP, India. ayush.20scse1010152@galgotiasuniversity.edu.in
Sovit Kumar Gupta, Student, Galgotias University Greater Noida, UP, India. sovit.20scse1010520@galgotiasuniversity.edu.in

The mobile application relies on the firebase database, which serves as a real-time cloud-based repository [8]. The trained model parameters are stored in the database, enabling real-time predictions. The research paper introduces several significant contributions, which are as follows:

1. The primary objective is to develop an efficient automated disease diagnosis model using machine learning techniques.
2. Three specific diseases, namely coronavirus, heart disease, and diabetes, are chosen for detailed analysis.
3. Logistic regression is employed for predicting computations.

1.2 Architecture

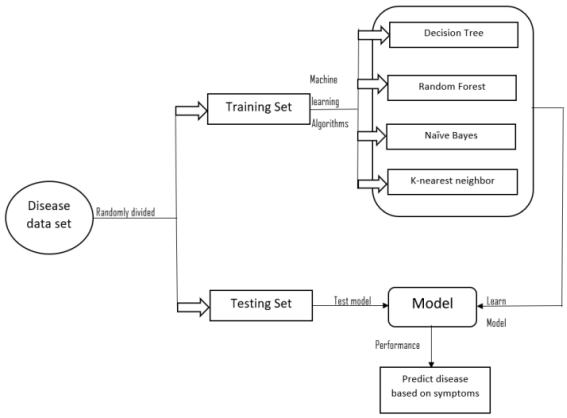

Fig.1 System Architecture

II. LITERATURE SURVEY

2.1 Overview

For this project, we conducted a thorough review of five papers from various sources to inform our research. Our focus was on examining the performance of different algorithms in various scenarios for disease prediction. The study's primary aim is to enhance treatment effectiveness by leveraging Machine Learning technology and simplifying the decision support system [2]. The researchers propose a comprehensive framework for heart disease diagnosis based on monitoring an individual's heartbeat [6]. This framework allows users to personalize their pulse requirements, and when a person's heartbeat exceeds the specified threshold, they receive a high pulse warning [4], indicating a potential risk of coronary failure or a heart attack. Through experiments utilizing a combination of different factors, the authors, Ahmed M. Alaa and Senthil Kumar Mohan, achieved an accuracy rate of 88.7% using a random hybrid forest model [9]. This paper focuses on the classic problem of supervised binary classification, where a dataset is provided with multiple attributes.

The dataset includes diverse features like plasma glucose concentration, blood pressure (mm Hg), body mass index, and age (years), sex, cholesterol, chest pain type etc [3].
Data mining techniques and machine learning techniques are employed in the healthcare sector to extract valuable information from patient datasets [6]. This paper specifically focuses on the prognosis of heart disease using supervised learning algorithms. The study utilizes Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Naïve Bayes algorithms to achieve this objective. The dataset used in the study comprises 3000 objects with 14 features[8]. While most of the existing literature analyzed a heart disease dataset with only 303 objects and 14 features, this paper aimed to build upon previous research by utilizing a larger dataset [10]. The findings reveal that Naive Bayes exhibited the highest performance, achieving an accuracy rate of 86.6%, [7]. demonstrating both high accuracy and computational efficiency [1]. In contrast, the Decision Tree algorithm obtained an accuracy of 78.69%, while the KNN algorithm achieved an accuracy of 77.85% [5][15][16].

2.2 Proposed System

The rapid expansion of big data in the biomedical and healthcare fields has yielded substantial advantages in precisely analyzing medical data. This analysis plays a pivotal role in early disease identification, enhancing patient care, and improving community services. Nevertheless, incomplete medical data compromises the accuracy of the analysis. Additionally, various regions possess distinct attributes associated with particular regional diseases, which can impede the accurate prediction of disease outbreaks. Our main aim is to improve machine learning algorithms to predict chronic disease outbreaks accurately in communities with a high prevalence of the disease. We conducted experiments using more modified prediction models and real hospital data collected from various regions within the county during the period from 2013 to 2015.

2.3 Problem Statement

Numerous studies have revealed that the majority of machine learning models developed for healthcare analysis focus only on a single disease. For instance, specific models are designed for analysing liver issues, cancer, and lung problems. [4] Consequently, individuals seeking accurate predictions across a broad spectrum of illnesses are required to consult multiple online resources [1]. The prediction of multiple diseases through a single analysis lacks a well-defined process. Inaccuracies in certain models can have severe implications for patient health [4]. Organizations aiming to evaluate their patients' medical records encounter additional time and financial investments due to the necessity of implementing diverse models.

2.4 Existing Systems

The current system focuses on forecasting particular chronic illnesses in a defined geographical area and community. It leverages Big Data and the CNN (Convolutional Neural Network) algorithm to estimate the risk of diseases [1]. For S type data, various Machine Learning algorithms are employed, such as K-nearest Neighbors, Decision Tree, and Naïve Bayesian. The system attains an impressive accuracy rate of up to 94.8% [7]. Researchers recently published a study where they applied ml techniques to enhance the prediction of chronic disease outbreaks in communities with a high disease prevalence [4]. The study involved conducting experiments using customized prediction models and utilizing real hospital data collected from some parts China [7]. The authors proposed a new algorithm called CNN-MDRP, which effectively integrates structured and unstructured data from the hospital [8].

III. METHODOLOGY

3.1 Methodology

Our project aims to develop a system for predicting multiple diseases based on symptoms entered by the patient. The project involves several tasks, starting with identifying the problem statement. We then prepare the dataset for analysis, ensuring its quality by examining anomalies, missing values, and other factors using techniques such as scatter plots and distribution graphs. The core aspect of our project revolves around employing Machine Learning techniques. Specifically, we utilize algorithms like SVM, Random Forest, and KNN to accurately predict diseases, enabling early detection and improved patient care. We are going to use many libraries of python which by name goes as:

```
import pandas as pd
import numpy as np
import plotly. graph objects as go
import plotly. express as px
import seaborn as sns
import matplotlib. pyplot as plt
```

In this dataset we have been provided with different information of a particular patient like their age, sex, cholesterol levels, resting bp, resting ecg etc. Our model analyses all these points and tell us whether the heart is at a 0 level or 1 level disease. 0 means healthy and 1 means affected bysome disease.

3.1.1 C-Support Vector Classifiers

SVMs are a popular type of supervised machine learning algorithm that is well-suited for both classification and regression tasks. The fundamental concept behind SVMs involves finding a hyperplane that maximizes the separation between different classes within the training data [7]. The determination of the hyperplane in SVM involves identifying the largest margin, which represents the distance between the hyperplane and the nearest data points from each class[5]. When the hyperplane is established, new data can be classified based on its position in relation to the hyperplane. SVMs are particularly valuable when dealing with high-dimensional data or when clear separation margins exist in the dataset [7]. SVM is a widely used algorithm for classification and regression, known for its straightforward implementation. While it is commonly used for classification, it can also be applied to regression problems.

3.1.2 Random Forest Classifiers

Random forest classifier is a widely employed supervised ml technique that finds its primary application in classification tasks, although it can also handle regression problems.

Retrieval Number: 100.1/ijaent.I96940812923
DOI: 10.35940/ijaent.I9694.11101123
Journal Website: www.ijaent.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

It falls under the category of ensemble learning methods. The implementation and utilization of Random Forest are straightforward, making it an ideal choice when time is limited for model development. The Random Forest algorithm creates numerous decision trees during the training phase and makes predictions by aggregating the results through a voting mechanism.

Data Flow Diagram

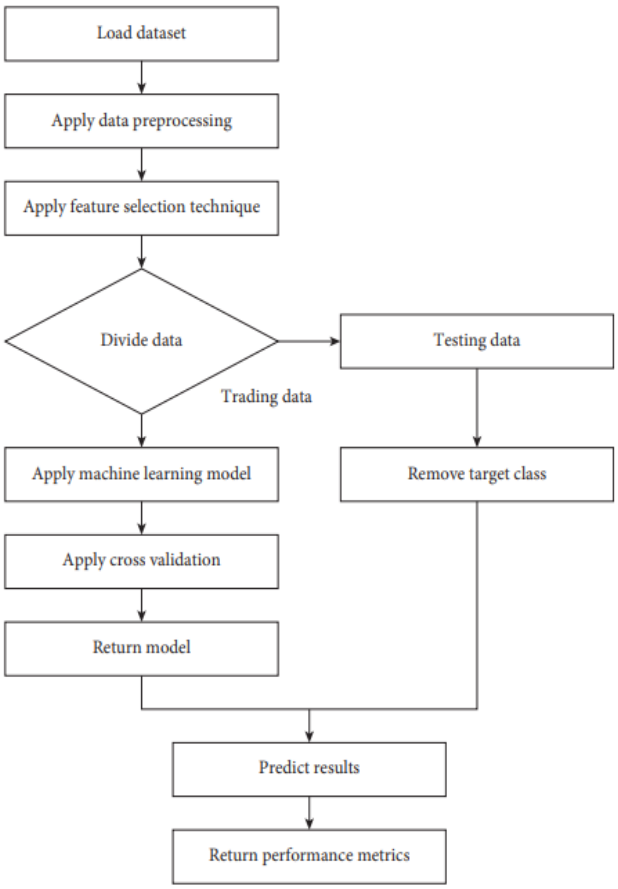

Fig. 3.1.1 Data Flow Diagram

3.1.3 KNN

The K-nearest neighbors (KNN) algorithm is a type of supervised learning method that classifies new data points by comparing their similarity to existing data. It is considered a non-parametric method as it does not assume any specific data distribution. KNN falls under the category of lazy learners because it does not undergo explicit training; instead, it retains the entire training dataset for classification purposes. KNN operates on the principle of feature similarity, where it assumes that similar objects exist in close proximity. It is an instance-based learning algorithm that approximates the local function.

IV. RESULT

The results of heart disease prediction using MLcan be highly accurate, with some studies reporting up to 80-90% accuracy. ML models are only as good as the data they are trained on, and there may be biasesor confounding factors that affect the predictions. Additionally, the predictions should be validated using independentdatasets and clinical trials before being applied in practice. After using all the machine learning algorithms, the result of the project is:

1. K-nearest Neighbors Classifiers

Accuracy	92.307692
Precision	91.176471
Recall	92.537313
F1-score	92.310711

2. C-support Vector Classifiers

Accuracy	89.510490
Precision	90.625000
Recall	86.567164
F1-score	89.491892

3. Random Forest Classifier

Accuracy	90.90909
Precision	86.48648
Recall	95.52238
F1-score	90.91531

Model Score Comparison

	K-Nearest Neighbors	C-Support Vector	Random Forest
Ac	92.307692	89.510490	90.90909
Pr	91.176471	90.625000	86.48648
Rc	92.537313	86.567164	95.52238
F1	92.310711	89.491892	90.91531

Keywords: Accuracy (Ac), Precision (Pr), Recall (Rc), F1-score (F1).

According to our focus the best recall score was present for the Random Forest classifier, but also this classifier presents the worst precision score, we must find a balance between recall score and precision score, for me the best model is the KNN classifier.

V. CONCLUSION

It is widely acknowledged that the advent of the Internet of Things and the digital age will bring forth new and fascinating technological possibilities in the field of medical treatment. In conclusion, ML has the potential to revolutionize heart disease prediction and improve patient outcomes. By leveraging the power of data and advanced algorithms, we can identify individuals at high risk of heart disease and provide targeted interventions to prevent or manage the condition. However, it is important to continue developing and refining these techniques, while also addressing the ethical and social implications of using AI in healthcare.

FUTURE SCOPE

There is potential for further research to expand the current study by exploring additional factors and elements. However, due to time constraints, the scope of this research is limited and future work is needed to delve deeper into these areas. There are plans to employ additional categorization methods, various discretization approaches, and several vote-by-classifier techniques. We propose including other diseases in the current system for future extension. Provide a chatbot to answer frequently asked questions and make the system as user-friendly as we can.

ACKNOWLEDGEMENT

We are grateful of the special guidance from our Guide and our Reviewer. We learned a lot about team working, leadership and some important technical aspects of our project. We thank our guide for their useful suggestions.

DECLARATION STATEMENT

Funding	No, I did not receive it.
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Not relevant.
Authors Contributions	All authors having equal contribution for this article.

REFERENCE

1. S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1– 16, 2019. https://doi.org/10.1186/s12911-019-1004-8
2. R. Katarya and P. Srinivas, "Predicting heart disease at early stages using machine learning: A survey," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302–305. https://doi.org/10.1109/ICESC48915.2020.9155586
3. Z. Ma, H. Ge, Y. Liu, M. Zhao, and J. Ma, "A combination method for android malware detection based on control flow graphs and machine learning algorithms," IEEE Access, vol. 7, pp. 21235–21245, 2019. https://doi.org/10.1109/ACCESS.2019.2896003
4. F. Tang, Y. Kawamoto, N. Kato, and J. Liu, "Future intelligent and secure vehicular network toward 6g: machine-learning approaches," Proceedings of the IEEE, vol. 108, no. 2, pp. 292–307, 2020. https://doi.org/10.1109/JPROC.2019.2954595
5. G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou, and D. Zhao, "Flight delay prediction based on aviation big data and machine learning," IEEE Transactions on Vehicular Technology, vol. 69, no. 1, pp. 140–150, 2020. https://doi.org/10.1109/TVT.2019.2954094
6. G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: wireless communication meets machine learning," IEEE Communications Magazine, vol. 58, no. 1, pp. 19– 25, 2020. https://doi.org/10.1109/MCOM.001.1900103
7. Konstantinos Psychogyios, Loukas Ilias, Christos Ntanos, Dimitris Askounis, "Missing Value Imputation Methods for Electronic Health Records", IEEE Access, vol.11, pp.21562-21574, 2023. https://doi.org/10.1109/ACCESS.2023.3251919
8. Shakil Ahmed Sumon, Raihan Goni, Niyaz Bin Hashem, Tanzil Shahria and Rashedur M. Rahman, "Violence Detection by Pretrained Modules with Different Deep Learning Approaches", *Vietnam Journal of Computer Science*, vol. 07, no. 01, pp. 19-40, 2020. https://doi.org/10.1142/S2196888820500013
9. Hammad, A. A. Abd El-Latif, A. Hussain et al., "Deep learning models for arrhythmia detection in IoT healthcare applications," *Computers and Electrical Engineering*, vol. 100, p. 108011, 2022. https://doi.org/10.1016/j.compeleceng.2022.108011
10. A. Sedik, M. Hammad, A. A. Abd El-Latif et al., "Deep learning modalities for biometric alteration detection in 5g networks-based secure smart cities," *IEEE Access*, vol. 9, pp. 94780–94788, 2021. https://doi.org/10.1109/ACCESS.2021.3088341
11. Saranya, Mrs. N., Kaviyarasu, P., Keerthana, A., & Oveya, C. (2020). Heart Disease Prediction using Machine Learning. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 9, Issue 1, pp. 700–704). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP. https://doi.org/10.35940/ijrte.f9780.059120 https://doi.org/10.35940/ijrte.F9780.059120
12. Reddy M, P. K., Reddy, T. S. K., Balakrishnan, S., Basha, S. M., & Poluru, R. K. (2019). Heart Disease Prediction Using Machine Learning Algorithm. In International Journal of Innovative Technology and Exploring Engineering (Vol. 8, Issue 10, pp. 2603–2606). https://doi.org/10.35940/ijitee.j9340.0881019
13. Prediction of Cardiovascular Disease using Machine Learning Algorithms. (2020). In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 3, pp. 2404–2414).. https://doi.org/10.35940/ijeat.b3986.029320
14. Harish, Mr. C. S., vamsi, Mr. G. gnana krishna, akhil, Mr. G. jaya phani, sravan, Mr. J. n v hari, & chowdary, Ms. V. mounika. (2021). Prediction of Heart Stroke using A Novel Framework – PySpark. In International Journal of Preventive Medicine and Health (Vol. 1, Issue 2, pp. 1–4). https://doi.org/10.54105/ijpmh.b1002.051221
15. D, Dr. K. (2021). VANET: Framework, Challenges and Applications. In Indian Journal of Data Communication and Networking (Vol. 1, Issue 2, pp. 10–15). https://doi.org/10.54105/ijdcn.b5002.041221
16. Varghese, A., Marri, M., & Chacko, Dr. S. (2023). Investigation of an Autonomous Vehicle's using Artificial Neural Network (ANN). In Indian Journal of Artificial Intelligence and Neural Networking (Vol. 3, Issue 6, pp. 1–11). https://doi.org/10.54105/ijainn.f1072.103623