International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Case Study on Early-Stage Risk Prediction by Machine Learning Algorithms

T. Deepthi[1], V.S. Triveni[2@], Thanniru Dharma Nithin[3], Danish Adnaan Mohammed[4], Salveru Keerthana[5]
1,2 Department of Mathematics, Geethanjali College of Engineering and Technology, Hyderabad, deepthitogercheti@gmail.com, vstriveni@gmail.com,
3,4,5 CSE-AIML Geethanjali College of Engineering and Technology, Hyderabad, dharmanithin29@gmail.com, danishadnaan518@gmail.com, salverukeerthana2003@gmail.com
@ Corresponding Author : V.S. Triveni, vstriveni@gmail.com

Abstract
In the healthcare sector, we generally require continuous monitoring of the patient, The model is designed to assist patients in understanding the potential hazards associated with the infectious illness by using Machine Learning algorithms. It provides recommendations as per the infection severity, enabling patients to effectively monitor their condition independently. All the data records are initialized in the dataset, these are stored in the database which helps in more accuracy of illness prediction. This software interface is simple, based on symptoms of the patient the algorithms will process and gives appropriate infection details, suggestions on severity of infection, recommends to consult a doctor or not. It also ensures the seasonal diseases ongoing in particular areas, identifies them, and gives regular alerts to the users. Non-Communicable Diseases (NCDs) are currently causing more infections are highlighted, and alerts the user by giving information and precautions to be taken for improvement of health. Predictive models and classification algorithms, examine the symptoms specified by the patient as input. Then the most possible disease name will be displayed as an output. Decision Tree, Naive Bayes Classifier, and Random Forest Algorithm are used to forecast the disease. Disease prediction is accomplished by employing ML Algorithms.

Keywords: Early-Stage Risk Prediction, Naive Bayes Classifier, Non-Communicable Diseases (NCDs), Random Forest Algorithm, Decision Tree Algorithm and Predictive models, Machine Learning (ML).

1. INTRODUCTION

The Disease Prediction framework that utilizes Machine Learning is specifically developed to forecast the occurrence of diseases. by leveraging user-provided information. It employs advanced algorithms such as Decision Trees, Naive Bayes, and Random Forest to analyze symptoms entered by the user, yielding precise and reliable results. This system not only assists users in identifying their ailments but also offers health maintenance tips. For users seeking information on non-serious conditions, it serves as a valuable tool by providing insights into the type of disease they may be experiencing. As the contemporary health industry assumes a pivotal role in treating patients, this system contributes significantly by offering an option available to individuals who choose not to physically visit hospitals or clinics. Through the input of symptoms and relevant information, users can ascertain the nature of their ailment. Health professionals can also benefit from this system by swiftly obtaining accurate information on a patient's condition.

The algorithms employed, namely Decision Trees, Naive Bayes, and Random Forest, collaboratively predict the most probable disease based on user-input symptoms. The system not only identifies potential diseases but also recommends precautionary measures to mitigate their progression. Additionally, it aids doctors in analyzing disease patterns within the broader societal context.

In this paper, we made an attempt to detect the disease prediction system, by employing data mining techniques; an initial analysis of the dataset is conducted. The main model is then trained using Machine Learning (ML) algorithms, which aids in predicting common diseases.

1.1 Problem Definition

Clinical decision-making often relies on doctors' intuition and experience rather than on a foundation of knowledge. This approach introduces unwanted biases, errors, and increased medical costs, to mitigate the negative impact on the quality of patient care, we are working on a unique program or project. This initiative leverages user-provided information and relevant datasets within the system to ensure precise and dependable outcomes.

In instances, where individuals are observing symptoms but are uncertain about the associated disease, the absence of clear understanding can potentially result in future health complications. To prevent this and enable timely

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

identification of diseases based on emerging symptoms, proactive measures are being taken. Our disease prediction initiative becomes a valuable resource. This tool is beneficial across diverse age groups, catering to children, teenagers, adults, and senior citizens alike, offering timely insights to improve overall healthcare outcomes.

1.2 Purpose of the Study

The aim of this study is to utilize both general information and reported symptoms to identify and forecast the particular disease(s) of a patient or group of patients. Implementation of this predictive system in the health industry has the potential to streamline the diagnostic process, reducing the workload on doctors and enabling them to efficiently anticipate and address patient diseases. The main moto of this prediction system is to provide information about different common illnesses. If these illnesses are not noticed or taken seriously, they can become more serious, creating big problems for the patient and their family. The system is good at guessing the most probable illness by looking at the symptoms, helping to detect it early and manage healthcare better.

1.3 Scope

Among various life-threatening diseases, one particular ailment has become a focal point of intensive medical research. Diagnosing this illness is a difficult task, leading us to investigate automated prediction methods to improve the effectiveness of future treatments. Several factors increase the chances of getting this disease, with smoking habits, cholesterol levels, familial medical history, and obesity. Healthcare organizations, including hospitals and medical centers, grapple with a significant challenge: delivering high-quality services while keeping costs reasonable. The essence of quality service lies in accurate patient diagnoses and the implementation of effective treatment plans. Addressing this challenge necessitates the extraction of patterns and relationships associated with the disease from historical databases. This method not only helps in automated predictions but also enables healthcare professionals to make well-informed decisions when dealing with complex diagnostic questions. The study's findings confirm that the suggested system has a unique potential in meeting the specified goals. Additionally, healthcare organizations encounter the added challenge of reducing costs related to clinical tests. This can be accomplished by strategically implementing computer-based information and decision support systems. Currently several hospitals use information systems for tasks like patient billing and inventory management, the integration of robust decision support systems is not widely practiced. Embracing advanced systems holds the promise of transforming healthcare operations, providing better diagnostic capabilities, and cost-effective solutions for enhanced patient care.

1.4 Features

The key features of the Case Study on Early-Stage Risk Prediction are outlined below:

1. Disease Prediction Based on Symptoms and General Information:

The project leverages datasets to predict diseases in patients by analyzing their symptoms and general information.

2. Utilization of Previous Hospital Datasets:

By comparing with historical hospital datasets, the project aims to provide accurate results, currently achieving up to 80% accuracy. Ongoing development endeavors are focused on achieving a goal of 100% accuracy.

3. Comprehensive Problem Solving and Prevention:

The Disease Prediction aspect of this study serves as a proactive tool, enabling the anticipation of diseases and addressing various health-related issues. It plays a crucial role in preventive healthcare.

4. Algorithmic Disease Prediction:

Diseases are predicted using advanced algorithms. Users input symptoms from a provided drop-down menu to ensure accuracy. It highlights the significance of providing detailed information about symptoms to ensure accurate results.

These features collectively position the project as a valuable tool in disease prediction, leveraging machine learning to enhance accuracy and contribute to proactive healthcare management.

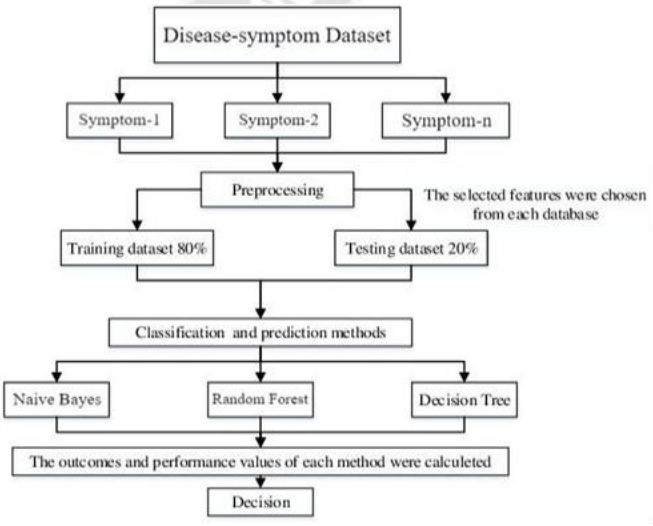

Fig 1 -Flow chart of Disease-Symptom

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

2. Literature Review

Md. Imam Hossain et.al. [5] explored the growing area of machine learning applications to identify critical risk factors and enhance diagnostic precision. Numerous studies have deployed algorithms such as Random Forest, Naïve Bayes, and Support Vector Machines (SVM) on datasets containing diverse patient attributes. Results accuracy varied by using random Forest method as 90% and logistic regression, neural networks, decision trees gave 73% to 99.2% accuracy due to diverse methodologies and datasets. The goal was to create intelligent systems for early detection and prognosis of heart disease, advancing healthcare with predictive technologies.

Rajneesh Thakur et al. [4] tried to predict diseases via symptom analysis, health information, and industry betterment using supervised algorithms in early stage. Notably, Random Forest method showed high precision at 0.96. The user-friendly interface allowed easy symptom input for accurate disease predictions.

Talasila Bhanuteja et al. [3] focused in implementing a machine learning-based framework for predicting diseases, analyzing clinical records of 4920 patients with 41 illnesses. Employing Decision Tree, Light GBM, and whereas Random Forest models gave accuracy results up to 98%. The models, selected based on provided indications, showcase the effectiveness of ML in early disease detection. The paper emphasized the evolvement of AI in data analysis, especially in healthcare, as technology enables the handling of vast datasets. The study concluded with insights into the potential of AI to contribute significantly to sophisticated data analysis in the future.

Chae et al. [6] focused on infectious diseases and the challenges they pose at individual and societal levels. The Korea Centre for Disease Control addressed limitations in its surveillance system, aiming to enhance response and prediction through deep learning algorithms and diverse datasets, including social media information. Improving fine-tuning the settings for Deep Neural Network (DNN) and Long Short-Term Memory (LSTM) models, the research forecasts infectious diseases one week ahead, outperforming the Autoregressive Integrated Moving Average (ARIMA) model. Results highlighted significant improvements in predicting chickenpox, with the top-10 DNN and LSTM models enhancing average performance by 24% and 19%, respectively.

Chen et al. [2] focused on predicting chronic diseases, including cerebral infarction, by integrating structured and unstructured healthcare data. Using a Chinese hospital dataset from 2013 to 2015, they employed a Convolutional Neural Network (CNN) for textual data feature extraction. The CNN-based multimodal disease risk prediction achieves an impressive 94.8% accuracy, outperforming traditional methods with faster convergence. The combined method provides valuable insights into assessing the risk of diseases, promising advanced predictive capabilities and contributing to more effective disease risk management in healthcare.

Feldman et al. [1] explored on personalized healthcare data tools, emphasizing challenges in scaling disease prediction algorithms and providing contextual information for clinical use. The paper addresses complexities in integrating tools within clinical time frames while ensuring interpretability for physicians by using the Collaborative Assessment and Recommendation Engine (CARE) algorithm. The findings offer insights for effective integration into clinical workflows, promoting patient-centric care.

Jyoti Soni et al. [7] did a comparative study on data mining techniques to predict clinical diagnosis of heart disease using methods like Decision Trees, k-nearest neighbor's (k-NN) and Naive Bayes algorithms, addressed biases in clinical decision-making and emphasized the data mining's potentiality in improving healthcare quality.

3. Methodology

The project utilizes several libraries, enhancing the functionality and capabilities of the Python programming language. Here's a summary of the libraries employed:

1. Python 3:

Python is a high-level, interpreted, interactive, and object-oriented scripting language. Known for its readability and simplicity, Python uses English keywords extensively, reducing syntactical complexities. It is designed to be user-friendly, making it suitable for a wide range of applications.

2. NumPy:

NumPy is a powerful library for Python, particularly designed for working with multi-dimensional arrays and matrices. It provides an extensive collection of high-level mathematical functions that operate seamlessly on these arrays. NumPy is commonly used for tasks involving array operations, mathematical functions, and numerical computations.

3. PyQt5:

PyQt5 is a GUI (Graphical User Interface) toolkit developed by Riverbank Computing. It serves as a Python interface for the Qt library, which is renowned for its versatility and cross-platform compatibility. PyQt5 combines the ease of Python programming with the robust features of Qt, enabling the development of interactive and visually appealing graphical user interfaces. It can also be integrated into C++ applications, allowing users to customize or improve the functionality of those applications.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

These libraries collectively add to the project's effectiveness by providing tools for array operations, mathematical computations, and the development of a user-friendly graphical interface.

3.1 Data Collection Processing

The dataset utilized in our project was carefully compiled from reliable sources, which notably include Kaggle and OpenAI. By tapping into Kaggle's extensive resources renowned for data science competitions, we accessed datasets specifically aligned with the scope of our project. Furthermore, we availed ourselves of the supplementary data provided by OpenAI, an influential entity in artificial intelligence research, thereby augmenting the richness and depth of our dataset.

To ensure the integrity our data, we attempted a meticulous data preprocessing phase. This phase encompassed tasks like addressing missing values, managing outliers, and standardizing data formats. Additionally, we implemented feature engineering techniques to extract pertinent information, thereby amplifying the predictive capabilities of our models. This thorough approach to data processing and collection establishes a solid foundation for robust and meaningful analyses.

The amalgamation of diverse datasets from Kaggle and OpenAI, coupled with deliberate preprocessing methods, empowers this paper with a comprehensive and top-tier dataset. Consequently, our study is better positioned to generate precise models and meaningful conclusions.

3.2 Results and Interpretation

Table 3.2.1 Accuracy Result on Algorithms

Algorithm	Accuracy before preprocessing	Accuracy after preprocessing
Gaussian Naive Bayes	88.08%	92.9%
Decision Tree	90.12%	93.85%
Random Forest	95.28%	97.64%

This paper has a primary focus on developing a predictive model for early-stage risk prediction, using a range of machine learning algorithms. The experimentation phase yielded results that underscored the efficacy of the algorithms both before and after preprocessing.

3.3 Interpretation:

3.3.1 Gaussian Naive Bayes:

The Gaussian Naive Bayes algorithm demonstrated a significant improvement in accuracy post-preprocessing, achieving an accuracy of 92.9%. This enhancement suggests that the data cleaning and feature engineering stages positively influenced the predictive capabilities of the Gaussian Naive Bayes algorithm.

3.3.2 Decision Tree:

The Decision Tree model exhibited a substantial increase in accuracy, rising from 90.12% to 93.85% after preprocessing. This highlights the pivotal role of data preparation in refining the model's performance and reliability.

3.3.3 Random Forest:

The Random Forest algorithm displayed impressive accuracy, starting at 95.28% and further improving to 97.64% after preprocessing. This emphasizes the robustness of the Random Forest model and underscores the impact of preprocessing steps on its predictive accuracy.

3.3.4 Key Insights:

Preprocessing emerged as a critical factor in enhancing the overall accuracy of all algorithms, emphasizing the pivotal role of data quality in machine learning model performance.

The Random Forest algorithm stood out as the most accurate model for early-stage risk prediction, demonstrating its effectiveness in handling complex relationships within the dataset.

These results and interpretations collectively highlight the success of the models in predicting early-stage risk and underscore the significance of meticulous data preprocessing in achieving optimal predictive performance.

Fig 2 illustrates that the Random Forest classifier outperforms the other two classifiers, making it the most suitable choice for the dataset. This conclusion further reinforces the effectiveness of the Random Forest algorithm in this context.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

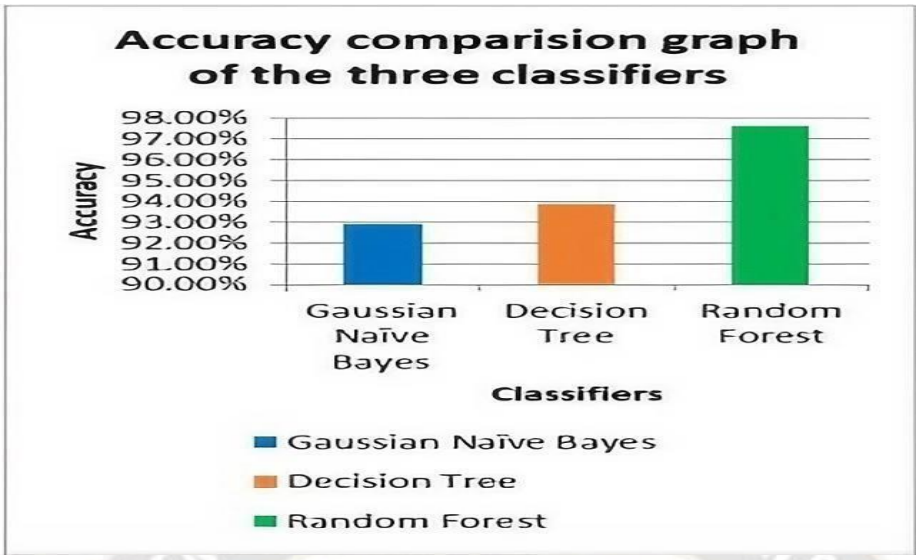

Fig 2- Accuracy Graph on classifiers

3.3.5 OUTPUT SCREEN

Manifestation form:

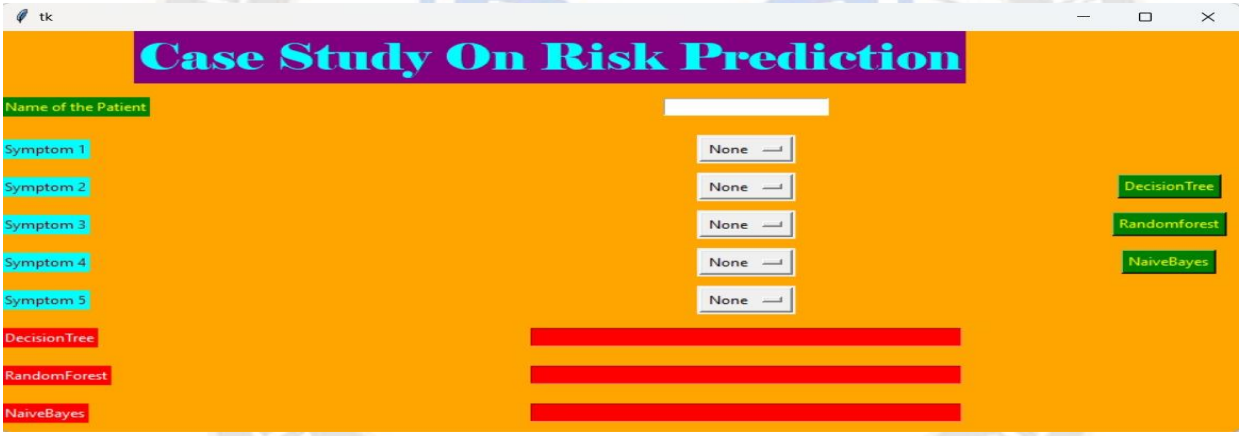

Fig 3 – Interface of user input

The predicted disease of the patient:

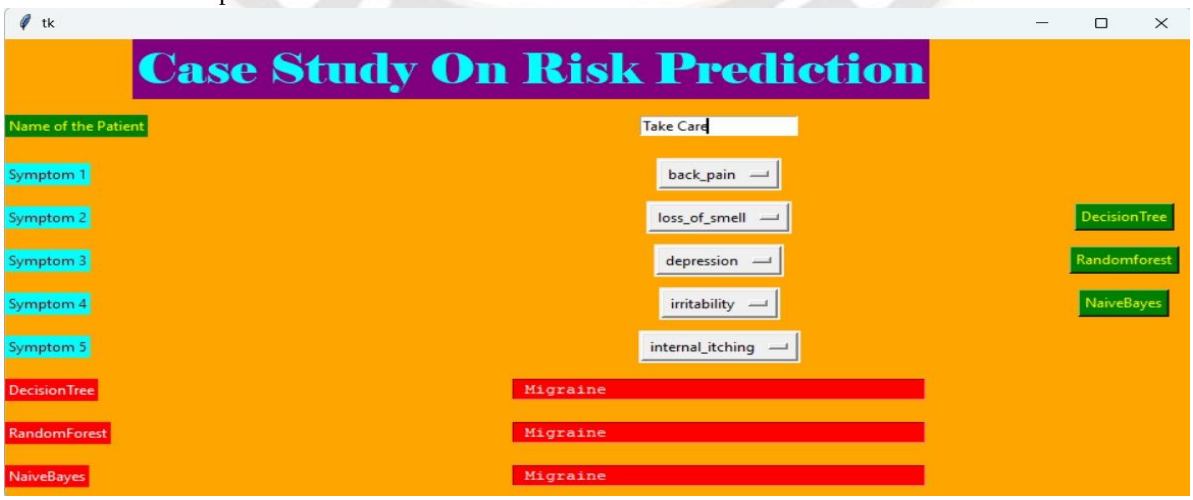

Fig 4 – Diagnosis Report

3.3.6 Conclusion

Summary stands out as a highly valuable and practical tool in the daily lives of individuals, especially within the healthcare sector. Its significance is particularly notable for healthcare professionals who rely on such systems to predict diseases based on patients' general information and symptoms. Given the crucial role played by the healthcare industry in patient treatment, this study emerges as a valuable asset. It offers an option for users who may choose not to visit hospitals or clinics, offering a user-friendly platform to understand their ailments by inputting symptoms and relevant information. This not only advantages users but also streamlines the workflow for healthcare professionals, potentially reducing the workload on doctors. The prototype of this prediction system incorporates three data mining classification modeling techniques, extracting hidden knowledge from a historical disease database. Significantly, the Naïve Bayes model, followed by Decision Tree and Random Forest, emerges as the most effective in predicting patients with diseases. Beyond prediction, the system enables efficient management of medicine resources required for treatment, leading to cost reduction and improved recovery processes. In essence, the successful implementation of this initiative holds the potential to revolutionize disease prediction, enhance healthcare delivery, and contribute to more efficient resource management in the healthcare sector. The broader impact on healthcare practices underscores the transformative power of integrating machine learning into the domain of disease prediction and management.

References:

[1] Keith Feldman, Darcy Davis, Nitesh V. Chawla (2015): “Scaling and contextualizing personalized healthcare: A Case Study of Disease Prediction Algorithm”, Journal of Biomedical Informatics, Volume 57, pp 377-385.

[2] M Chen et al. (2017):” Disease Prediction by ML over Big Data from Healthcare Communities.” IEEE Access, Volume 5, pp 8869-8879.

[3] Talasila Bhanuteja et al. (2021): “Symptoms Based Multiple Disease Prediction Model using ML Approach” International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-10 Issue-9, pp 67-72.

[4] Rajneesh Thakur, Mansha, Pranjal Sharma, Dhruv (2023): "Disease Prediction using Classification Algorithm." International Journal for Research in Applied Science and Engineering Technology (IJRASET) Volume 11, Issue 8, pp 675-682.

[5] Md. Imam Hossain et al. (2023): "Heart disease prediction using distinct artificial intelligence techniques: performance analysis and comparison", Iron Journal of Computer Science, Volume 6, pp 397-417.

[6] S Chae, S Kwon, D Lee (2018): “Predicting Infectious Disease Using Deep Learning and Big Data”, International journal of environment research and public health, Volume 15, Issue 8.

[7] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni (2011): "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, International Journal of Computer Applications, Volume 17- No.8, pp 43-48.