International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Machine Learning Techniques Based Hybrid Model for Disease Prediction

Neha Kukade[1], Sunil Gupta[2]

[1,2]Department of Computer Science and Engineering
[1,2]Prof. Ram Meghe Institute of Technology and Research, Badnera – Amravati 444701
[1]nehaskukade@gmail.com, [2]srgupta@mitra.ac.in

Abstract

In the era of digital healthcare, machine learning plays a crucial role in the early prediction and diagnosis of diseases, enabling timely medical intervention. This paper presents an analytical study of various machine learning techniques used for instant disease prediction based on patient symptoms. The performance of algorithms such as Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN) is evaluated using medical datasets containing symptom-disease pairs. These algorithms are compared in terms of prediction accuracy, processing time, and reliability. The system accepts symptom input from users and predicts the most probable disease, thereby assisting patients in identifying potential health issues before consulting a medical professional. The findings aim to identify the most effective algorithm for real-time disease diagnosis, contributing to the development of intelligent and accessible healthcare systems.

Keywords— Machine Learning, Disease Prediction, Decision Tree, Random Forest, Naive Bayes, KNN, Healthcare, Medical Diagnosis.

I. INTRODUCTION

In recent years, the application of machine learning in the healthcare domain has witnessed significant growth, particularly in the area of disease prediction and diagnosis. One such advancement is the development of web-based disease prediction systems that offer preliminary guidance to users based on their symptoms. These systems aim to predict the most probable disease by analysing user-reported symptoms using machine learning algorithms and data sets collected from various health-related sources.

With the increasing accessibility of the internet, many individuals seek online solutions for health concerns before visiting healthcare facilities. The convenience of consulting such systems from one's home makes them highly valuable, especially in areas with limited access to medical professionals. By leveraging data processing methods and machine learning techniques, these platforms can provide accurate disease predictions based on the user's input. Consequently, the system's final output enables patients to connect with relevant healthcare specialists for further treatment, thereby streamlining the initial phase of medical consultation.

This paper presents an analytical study of various machine learning techniques—such as Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes—that are employed for instant disease prediction and diagnosis. The study evaluates the performance, accuracy, and suitability of these models in real-world healthcare applications.

II. PROBLEM ANALYSIS

A. Existing System

In prior studies, disease prediction systems have primarily focused on diagnosing specific conditions only. For instance, CNN algorithms have been applied to forecast disease risks. Existing systems commonly use supervised learning techniques such as Decision Trees, Naive Bayes, and Random Forest classifiers. These models, when trained on clinical datasets, have achieved prediction accuracies of up to 84% [2]. However, these systems are often limited to structured hospital data from predefined patient records and target only a narrow scope of illnesses. With the evolution of medical datasets that now include both structured and unstructured data formats, there is a growing need for more flexible and generalized disease prediction frameworks. To address this, researchers have introduced approaches like the multimodal disease risk prediction algorithm, incorporating methods such as CNNs [3].

B. Proposed System

The proposed model is developed to enable automatic disease prediction using machine learning framework trained on comprehensive health-related datasets. This system not only predicts the probable diseases based on symptoms but also provides a confidence score reflecting the certainty of each prediction. When a likely disease is detected, the system recommends a relevant healthcare professional for online consultation, aiding in early intervention [2]. Designed as a clinical decision support tool, the system aims to assist healthcare providers by delivering diagnostic suggestions and reinforcing medical decisions. The primary objective of this work is to develop a web-based platform capable of predicting diseases from a broad range of symptoms with probabilistic outputs, thereby supporting timely diagnosis and improving healthcare outcomes [3].

C. Related Works

1. Data Collection

To enable accurate identification of disease symptoms, structured symptom data has been made available online. This ensures a higher degree of precision, as the dataset does not contain synthetic or placeholder entries. For this study, a publicly accessible dataset was obtained from Kaggle. The dataset is in CSV format and comprises 5000 records and 133 attributes. Out of these, 132 attributes represent distinct symptoms, while the final column specifies the disease class. There are 40 unique disease classes identified within the dataset [5].

Group 1: Swine Flu Related	**Group 2: Paralysis and Related Conditions**	**Group 3: Ebola and Associated Disorders**
Heart-related Conditions	Migraine	Respiratory Infections
Liver Disorders (Jaundice)	Degenerative Joint Disease (Osteoarthritis)	Acquired Immunodeficiency Syndrome (AIDS)
Autoimmune Skin Disorder	Hepatitis B	Contagious Skin Infection (Impetigo)

Group 1: Swine Flu Related	Group 2: Paralysis and Related Conditions	Group 3: Ebola and Associated Disorders
Inflammatory Joint Condition	Hepatitis C	Psoriasis
Hemorrhoidal Disease	Recurring Hepatitis B	Vascular Pressure Imbalance (Hypertension)
Blood Sugar Disorders (Diabetes)	Alcohol-Induced Liver Disease	Hepatitis D
Thyroid Hormone Imbalance	Adverse Drug Response	Hepatitis A
Hepatitis E	Urinary System Infection	Skin Condition (Acne)
Digestive Tract Ulcer	Tuberculosis	Liver Function Disorder (Chronic Cholestasis)
Gastroesophageal Reflux Disease	Mosquito-borne Infection (Dengue)	Blood Sugar Deficiency (Hypoglycaemia)
Respiratory Disorder (Asthma)	Spinal Degeneration (Cervical Spondylosis)	Vertigo
Cervical/Spinal Spondylosis	Dizziness/Vertigo	Viral Skin Infection (Chickenpox)
Chickenpox	Gastrointestinal Inflammation	Underactive Thyroid (Hypothyroidism)
Allergy Response	Recurrent Chickenpox	Endocrine Disorder (Hypothyroidism)
Lower Back Pain	Pneumonia	Chickenpox

Table 1. Categorized List of Diseases Used in Dataset [8]

2. Data Cleaning

All instances of missing data within the dataset were addressed during the preprocessing phase. In this study, no null or missing values were present. To maintain data quality and integrity, commonly adopted data cleansing methods, including distance-based and clustering-based outlier detection techniques, are applied when necessary [8].

3. Data Reduction

In order to improve algorithmic performance on datasets with lower dimensions, data from higher dimensions is transformed by eliminating redundant features. This dimensionality reduction process enhances efficiency, particularly when working with high-dimensional clinical data. Redundant symptoms

are identified and removed using correlation coefficients (denoted as r), which helps in selecting the most informative features. Out of 132 original symptoms [8], a total of 90 were retained for further analysis after the removal of highly correlated features.

The correlation coefficient r is computed as follows:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

r: Pearson correlation coefficient

xi: Value of the x variable

$\bar{x}$: Mean of the x variable

yi: Value of the y variable

$\bar{y}$: Mean of the y variable

III. METHODOLOGY AND APPLIED ALGORITHMS

A. Overview

This study employs supervised machine learning algorithms for the development of a disease prediction system. Classification techniques were selected due to their efficacy in predictive modelling tasks. Specifically, three widely recognized classification algorithms were utilized: Decision Tree, Random Forest, and Naïve Bayes. The Naïve Bayes classifier was further evaluated using its three standard variants: Gaussian, Multinomial, and Bernoulli.

B. Decision Tree Classifier

The Decision Tree algorithm is a robust and interpretable model frequently applied in classification and pattern recognition tasks. Its hierarchical structure includes a root node, internal nodes, and leaf nodes. The root node represents the attribute with the highest information gain, while internal nodes perform conditional splits to derive outcomes represented by the leaves.

In this work, the Gain Ratio criterion was employed, which uses entropy-based measures to determine attribute selection. The Information Gain (IG) is defined as:

$IG(C, A) = E(C) - E(C|A)$

where E(C) denotes the entropy of the class and E(C|A) is the entropy conditioned on attribute A. The Decision Tree classifier achieved an accuracy of approximately 95%, indicating strong performance in disease classification scenarios [1] – [5].

C. Random Forest Classifier

Random Forest is an ensemble-based algorithm that constructs a multitude of decision trees during training. The final prediction is obtained via majority voting across all trees, thereby improving generalization and reducing overfitting.

The classification procedure involves the following steps:

1. Random feature subsets are selected to train each decision tree.
2. Multiple trees are generated using bootstrapped datasets.
3. Each tree provides a prediction, and the most frequent class label is selected as the final output.

This method is particularly effective for datasets with high dimensionality and complex feature interactions [1] – [5].

D. Naïve Bayes Classifier

The Naïve Bayes classifier applies Bayes' Theorem under the assumption of feature independence. It calculates the posterior probability of a class given input features as:

$$P(Y \mid X_1, X_2, \ldots, X_n) = \frac{P(X_1, X_2, \ldots, X_n \mid Y) \cdot P(Y)}{P(X_1, X_2, \ldots, X_n)}$$

where Y represents the class label and X1, X2, ..., Xn are the features.

Three models of Naïve Bayes were considered:

- **Gaussian Naïve Bayes**, which assumes continuous features follow a normal distribution;
- **Multinomial Naïve Bayes**, suitable for discrete counts;
- **Bernoulli Naïve Bayes**, designed for binary-valued features.

For this implementation, the Gaussian Naïve Bayes model was adopted using the GaussianNB() function from the scikit-learn library due to its simplicity and effectiveness with continuous input data [5] – [7].

IV: IMPLEMENTATION DETAILS

1. Disease Prediction with Integrated Online Consultation

To enhance accessibility and provide timely healthcare services, the proposed system is designed with an intuitive web-based interface comprising two primary user roles:

- **Patients**
- **Doctors**

Each role is equipped with a tailored dashboard to facilitate relevant functionalities. The system aims to bridge the gap between symptom identification and professional consultation by integrating machine learning-based disease prediction

with a direct communication channel between doctors and patients.

2. Patient Dashboard Functionalities:

- **Symptom-based Disease Prediction**: Patients can input symptoms, which are analyzed using trained machine learning models such as Naive Bayes, Decision Tree, K-Nearest Neighbors (KNN), and Random Forest to predict probable diseases.
- **Online Consultation**: Patients can initiate consultations with registered doctors through a secure interface, enabling real-time interaction for advice and treatment.
- **Feedback Mechanism**: After consultation, patients can provide feedback to improve service quality and track satisfaction.
- **Profile Management**: Patients can view and update their personal and medical profile details as needed.

3. Doctor Dashboard Functionalities:

- **Consultation Interface**: Doctors can view consultation requests, access patient records, and provide diagnoses or recommendations through the system.
- **Feedback Access**: Doctors can review feedback to evaluate performance and improve interaction quality.
- **Profile Update**: Doctors are allowed to modify their credentials and availability status.

This system ensures a direct, secure, and efficient communication channel between patients and healthcare professionals, explicitly excluding third-party intermediaries to maintain confidentiality and streamline interactions.

The inclusion of disease prediction algorithms enhances the decision-making process by providing an initial automated assessment, which doctors can verify and build upon during the consultation process. This integration supports instant preliminary diagnosis and patient care, addressing delays associated with traditional healthcare systems.

- **Web-Based Application for Instant Disease Prediction and Online Consultation**

The primary objective of the proposed system is to develop a robust web-based application capable of predicting diseases based on user-reported symptoms through the integration of machine learning algorithms. The system architecture supports three categories of authenticated users: patients, healthcare professionals (doctors), and system administrators. Each user category is granted access to specific functionalities tailored to their role within the platform.

Patients interact with the system by submitting their symptoms through an intuitive user interface. The system processes the input data and applies trained machine learning models such as Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN) to identify the most likely disease or health condition associated with the reported symptoms. The system predicts the disease along with an associated confidence score, which reflects the model's prediction accuracy.

Following the prediction, the system recommends relevant medical professionals for online consultation. To access the platform, users must register and log in securely. Once authenticated, patients can utilize a range of features, including symptom-based diagnosis, access to consultation services, and the ability to provide feedback. Additionally, patients can update their profile and view consultation history, ensuring continuity of care.

For healthcare professionals, the system enables access to patient-submitted symptom data and predicted diagnoses. Doctors can view, respond to, and manage consultations, thereby delivering prompt and informed medical advice. They can also update their profiles and review patient feedback.

The system facilitates seamless communication between patients and doctors without involving third-party intermediaries. The platform is built with a user-centric approach, ensuring simplicity, responsiveness, and reliability. The front-end interface is developed using HTML5, CSS, JavaScript, and jQuery, while Django is used for the back-end framework. PostgreSQL serves as the primary database, with PgAdmin used for database administration.

This integrated framework ensures efficient diagnosis, immediate consultation, and personalized medical support, thus bridging the gap between symptom recognition and timely healthcare delivery.

V. RESULTS AND DISCUSSION

The proposed model was trained on a dataset comprising 5,000 records and 132 unique symptom attributes. To mitigate the risk of overfitting and improve model generalization, dimensionality reduction was applied by selecting the most relevant 90 features from the original 132 symptom variables. This optimized feature set was then used to evaluate the performance of three distinct machine learning algorithms: Naïve Bayes, Decision Tree, and Random Forest.

Among the evaluated classifiers, the Naïve Bayes algorithm demonstrated superior performance, achieving the highest training accuracy of 94.71%, as summarized in Table I. This indicates the algorithm's strong capability in handling the probabilistic classification of symptoms and diseases. In comparison, both the Decision Tree and Random Forest algorithms also yielded high accuracy but were slightly outperformed by Naïve Bayes in terms of overall efficiency and consistency.

These findings highlight the effectiveness of probabilistic models like Naïve Bayes in instant disease diagnosis tasks, where rapid and accurate predictions are crucial for timely patient care and medical decision-making.

ALGORITHM USED	ACCURACY SCORE
Decision Tree	0.9763
Random Forest	0.9763
Naïve Bayes	0.9812

Table : Training Accuracy

All the evaluated machine learning techniques were initially trained using historical patient data. Subsequently, the models were validated using previously unseen symptom records to assess their ability to generalize to new illnesses. Table II presents the accuracy scores of the algorithms, along with the number of diseases correctly and incorrectly classified. Among the evaluated models, the Naïve Bayes classifier achieved the highest accuracy of 97%, demonstrating its robustness and reliability in predicting disease outcomes [8].

ALGORITHM	ACCURACY SCORE	CORRECTLY CLASSIFIED	INCORRECTLY CLASSIFIED
Decision Tree	0.94	38	2
Random Forest	0.94	38	2
Naïve Bayes	0.96	40	1

Table.3 Accuracy and confusion Matrix

VI. CONCLUSION

A web-based disease prediction system utilizing machine learning techniques was successfully developed to support instant diagnosis and patient care. Decision Tree, Naïve Bayes, and Random Forest algorithms were employed to build and compare predictive models, which were further integrated to enhance diagnostic accuracy. Given the rapid growth of medical data, efficient data processing is critical to accurately identify potential diseases based on symptom inputs. By leveraging patient symptom data, the system was able to generate reliable risk predictions for general illnesses. Additionally, the platform facilitates remote consultations by allowing patients to interact directly with healthcare professionals. This framework not only offers a practical and scalable solution but also demonstrates high prediction accuracy and fairness in its results.

REFERENCE:

1. S. Kulkarni, I. Sawant, M. Shinde, and V. Sonar, “Django website for disease prediction using machine learning,” J. Univ. Shanghai Sci. Technol., Jun. 2021.
2. Naveenkumar, Kirubhakaran, Jeeva, Shobana, and Sangeetha, “Smart health prediction using machine learning,” Special Issue of 2nd Int. Conf. Adv. Res. Dev. (ICARD), 2021.
3. K. Pingale, S. Surwase, V. Kulkarni, S. Sarage, and A. Karve, “Disease prediction using machine learning,” Int. Res. J. Eng. Technol. (IRJET), vol. 6, no. 12, Dec. 2019.
4. Y. Tambe, S. Awhad, A. Keny, and M. Kulkarni, “Disease predictions using machine learning and online consultations,” J. Univ. Shanghai Sci. Technol., Jun. 2021.
5. N. Yede, R. Koul, C. Harde, K. Gaurav, and C. S. Pagar, “General disease prediction based on symptoms provided by patient,” Open Access Int. J. Sci. Eng. Technol., Jun. 2021.
6. P. Panapana, K. S. R. Reddy, J. Deepika, G. R. Babu, and A. Drakshayani, “Disease prediction and medication advice using machine learning algorithms,” Int. J. Creative Res. Thoughts (IJCRT), vol. 11, no. 3, Mar. 2023.
7. T. R. Patil and S. S. Sherekar, “Performance analysis of Naive Bayes and J48 classification algorithm for data classification,” Int. J. Comput. Sci. Appl., vol. 6, no. 2, Apr. 2013.
8. J. P. Gupta, A. Singh, and R. K. Kumar, “A computer-based disease prediction and medicine recommendation system using machine learning approach,” Int. J. Adv. Res. Eng. Technol. (IJARET), vol. 12, no. 3, Mar. 2021.
9. P. Sharma and S. Sharma, “A comprehensive study of the machine learning with federated learning approach for predicting heart disease,” in Proc. 2023 6th Int. Conf. Contemporary Comput. Informat. (IC3I), Ghaziabad, India, 2023.
10. R. Mukhlis, J. Hussain, S. Saleem, A. A. Aydin, H. Kwon, and M. A. Al-antari, “A novel CAD framework with visual and textual interpretability: Multimodal insights for predicting respiratory diseases,” in Proc. 8th Int. Artif. Intell. Data Process. Symp. (IDAP), Malatya, Türkiye, Sep. 2024.

11. N. R. Barman, K. Sharma, and R. Hazra, “A transformer-based approach to automate disease prediction from patient descriptions,” in Proc. 2023 IEEE 7th Conf. Inf. Commun. Technol. (CICT), Silchar, India, 2023.
12. R. Zhao, S. Chen, J. Liu, and J. Zhang, “Confidence-guided weakly-supervised visual evidence discovering for trustworthy glaucoma diagnosis,” in Proc. 2023 IEEE Int. Conf. Bioinformatics Biomed. (BIBM), Changsha, China, 2023, pp. 2416–2423.
13. H. S. Hemalatha, R. Thangarajan, S. Selvaraj, P. Dhanya Sree, P. Malathi, and V. Vignesh, “Disease prediction and specialist recommendation using machine learning,” in Proc. 2024 1st Int. Conf. Adv. Comput., Commun. Netw. (ICAC2N), Erode, India, 2024, pp. 1081–1086.
14. P. Jampani, S. V. Tanuku, V. R. Javvadi, A. K. Amarendra, V. S. Nandi, and B. S. N. Benarji, “Exploring Alzheimer's disease risk prediction: A data-driven approach with RandomForestClassifier,” in Proc. 2024 8th Int. Conf. Inventive Syst. Control (ICISC), Vaddeswaram, India, 2024.
15. P. Parasdeep and L. Chand, “Implementation of disease prediction algorithm using machine learning,” Int. J. Eng. Res. Comput. Sci. Eng. (IJERCSE), vol. 11, no. 8, pp. 29–33, Aug. 2024.