Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Intelligent Disease Prediction and Personalized Health Management System

Shanthi Makka[1], Ramesh Babu Popuri[2], Rajendar Sandiri[3], Muni Sekhar Velpuru[4], Gagandeep Arora[5]

[1]Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, Telangana, India, dr.shanthimakka@gmail.com
[2]Malineni Perumallu Educational Society's Group of Institutions, Pulladigunta, Guntur, Andhra Pradesh, India, popuri.ramesh2006@gmail.com
[3]Department of Electronics and Communication Engineering, Vardhaman College of Engineering, Hyderabad, India, sandiri.rajendar@gmail.com
[4]Department of Information Technology, Vardhaman College of Engineering, Hyderabad, Telangana, India, Munisek@gmail.com
[5]Department of AIML, Vardhaman College of Engineering, Hyderabad, India, Gagandeeparora250379@gmail.com

An increasing number of effective and precise illness prediction techniques are required due to the volume of healthcare data that is being collected. The application of machine learning techniques to the creation of an intelligent disease prediction system is examined in this research. The suggested model makes use of sophisticated algorithms to examine various patient data sets, such as those pertaining to lifestyle, medical history, and demographics. In order to create a prediction model that can identify possible health hazards, the paper focuses on integrating ensemble learning with machine learning techniques such as decision trees, random forest and etc. The model is trained and evaluated on a large dataset that is representative of a range of diseases and the factors that influence them. The findings show how machine learning may be used for risk assessment and early disease identification, giving healthcare professionals a useful tool for wise decision-making and efficient resource allocation. By integrating intelligent illness prediction systems, the research findings support ongoing efforts to improve healthcare outcomes, lower treatment costs, and increase overall patient well-being.

Keywords: Machine Learning; Prediction; Disease; Decision Trees; Ensemble Learning.

1. Introduction

Integration of cutting-edge technologies has become essential for early disease identification

and efficient disease management in the constantly changing field of healthcare. As a branch of artificial intelligence, machine learning has become a potent instrument that can scan enormous volumes of data and identify patterns, trends, and correlations that may be missed by more conventional diagnostic techniques. This paper explores the field of intelligent disease prediction through machine learning, with the goal of determining how advanced algorithms can transform the diagnosis and early identification of a range of medical diseases. It is possible to overestimate the significance of predictive analytics as there are few difficulties brought on by the growing complexity of diseases. With the help of large datasets including genetic data, medical records Wickramasinghe et al. (2021) and lifestyle factors, machine-learning algorithms may be able to predict the beginning of diseases and tailor treatments for the best possible outcomes for patients. This work aims to clarify the state of intelligent disease prediction today, highlighting the approaches, obstacles, and innovations that characterize this quickly developing area Makka, S. et al. (2021). This paper emphasizes the value of using machine learning for real-time epidemic surveillance and provides insight into how these innovations can improve our capacity to track, anticipate, and treat infectious diseases with previously unheard-of precision and speed. In parallel, explored how predictive modeling might be used to forecast the beginning and course of chronic illnesses, providing tailored insights for more efficient and focused interventions. Global healthcare systems must contend with the dual challenges of an increasing amount of patient data and the need to deliver accurate and fast diagnoses. Conventional diagnostic techniques, however beneficial, frequently struggle with the intricacy and vast amount of accessible medical data Sunitha et al. (2022). A strong answer to this problem is provided by the development of machine learning, which is distinguished by its capacity to identify patterns and extract insights from enormous datasets. Using emerging technologies like AI and machine learning can used to develop personalized healthcare management system.

2. Literature Survey

Many researches have been done to develop an application for predicting different types of diseases. Because of rising of different types of disease is one of the global issues in the world; so many researchers have worked on different types of solutions to predict disease with the help of symptoms. Many researchers have worked on how to use latest technology like chatGPT for training of better model so that it predicts the disease with more accuracy Jungwirth et al. (2023). Some researchers worked on using of advanced algorithms such as LDA (Linear Discriminant Analysis) and ASV-RF (Advanced-Spatial-Vector-based Random Forest) for training of best model Menonu P., et al. (2023). Using of AI in forecasting of epidemic/pandemic diseases also helped researchers to find a solution for predicting diseases by taking symptoms as input Bhattamisra et al. (2023), Meenigea et al. (2023). Emerging technologies like HDT (Human Digital Twin) was also used by some of the researchers in-order to find solution for personalized healthcare management systems by using of AI Okegbile, Samuel D., et al. (2022).

Over the past decade, information analysis has significantly expanded, driven by explicit rule-based algorithms termed "AI." AI combines with data analysis to create predictive models, utilizing predefined rules to analyze data and forecast outcomes Sivabalaselvamani, D., et al.

(2021). Semantic networks were also used by some of the researchers to train a best predictive model. Semantic networks are also able to train a model for prediction Kobrinskii, B. A., et al. (2019). In order to develop a smart healthcare system, some researchers also tried to use IoT in their applications so that every task will be automated. IoT can be used in almost every area; because of huge development in IoT every thing became smart Tian, Shuo, et al. (2019). Evolution of predictive algorithms made tasks simpler in the of machine learning. Many researchers used these predictive algorithms for predicting the disease by taking symptoms as input Grampurohit et al. (2020, Makka, S. et al. (2022). Lot of researchers worked on many ensemble-learning techniques so that can able to train best predictive model Uddin, Shahadat, et al. (2019).

Random Forest is one of the examples for ensemble learning technique under the category of bagging. During training, the machine-learning ensemble, Random Forest learning algorithm generates numerous decision trees. By combining the results of several trees, it reduces over fitting and increases robustness in classification and regression problems, boosting predictive accuracy and generalization Shah et al., (2020), Pingale, Kedar, et al. (2019). Gradient Boosting entails assembling weak learners, typically decision trees, to construct a predictive model step-by-step. It highlights cases that were incorrectly classified while gradually fixing the flaws in the prior model. In regression and classification problems, gradient boosting, as demonstrated by algorithms such as XGBoost, frequently yields high-predicted accuracy Patel et al. (2015). In machine learning, decision trees are models that resemble trees and are used for regression and classification. They produced a structure of judgments by recursively dividing the data according to features. Decision trees facilitate decision-making processes by associating input features with output predictions in a straightforward, comprehensible, and efficient manner Biau et al., (2016), Natekin et al. (2013), Myles, Anthony J., et al. (2004).

The probabilistic classification algorithm Naive Bayes relies on the principles of Bayes' theorem. By assuming feature independence, computations are made simpler. Naive Bayes is a popular text classification and spam-filtering algorithm that determines the most likely class for newly observed occurrences by utilizing observed attributes and previous probability Song et al. (2015, Priyam, Anuja, et al. (2013).

3. Proposed Model

3.1 Methodology

a. Data Collection

Obtained health information from a several of sources, covering an extensive range of patient characteristics, such as demographics, Makka, S. et al. (2022) history, lifestyle choices, and relevant biomarkers.

b. Data Preprocessing

Carried out thorough data preprocessing to guarantee the dataset's quality and integrity. This required dealing with outliers, missing values, and standardizing or normalizing numerical characteristics. The categorical variables were appropriately encoded to ensure compatibility with machine learning algorithms.

c. Algorithm Exploration

Explored a variety of machine Learning algorithms, which include methodologies like decision trees, neural networks, random forests, and support vector machines. This step aimed to identify the algorithms most suited for the intelligent prediction of diseases.

d. Accuracy Comparison

Evaluated several algorithms performances quantitatively using important metrics such as precision, F1 score, recall, accuracy, and area under the receiver operating characteristic curve (AUC-ROC). Carried out a comprehensive comparative investigation to identify each algorithm's advantages and disadvantages.

e. Algorithm Selection

Choose the best algorithm for illness prediction based on the thorough comparison. High accuracy, robustness, and conformity with the project's goals were among the selection criteria.

f. Accuracy Verification

Tested the resulting algorithm thoroughly on separate datasets to confirm its accuracy. Make sure the selected algorithm performed consistently and dependably in a range of situations.

g. Result Visualization

Using powerful visualization tools, the results of the intelligent illness prediction model were presented. To enhance the readability of the data and provide insights into the prediction power of the model, graphical displays such as charts and plots were used

3.2 Proposed Architecture

An architectural diagram visually represents a system's components, connections, and interactions. It offers a concise overview of how various parts of the system are structured and how they interact. The diagram's interfaces highlight the interaction points between different functionalities or modules, while relationships

show how components work together to meet system requirements, indicating dependencies and connections. The architecture also includes layers representing different levels of abstraction, such as presentation, application logic, and data layers. Furthermore, it may illustrate technologies, deployment methods, and architectural constraints. As a communication tool, architectural diagrams help stakeholders understand the system's architecture, facilitating discussions about development, scalability, performance, security, and compliance. In summary, it provides a visually organized and intuitive view of a system's architecture, enabling effective decision-making and collaboration throughout the software development lifecycle.

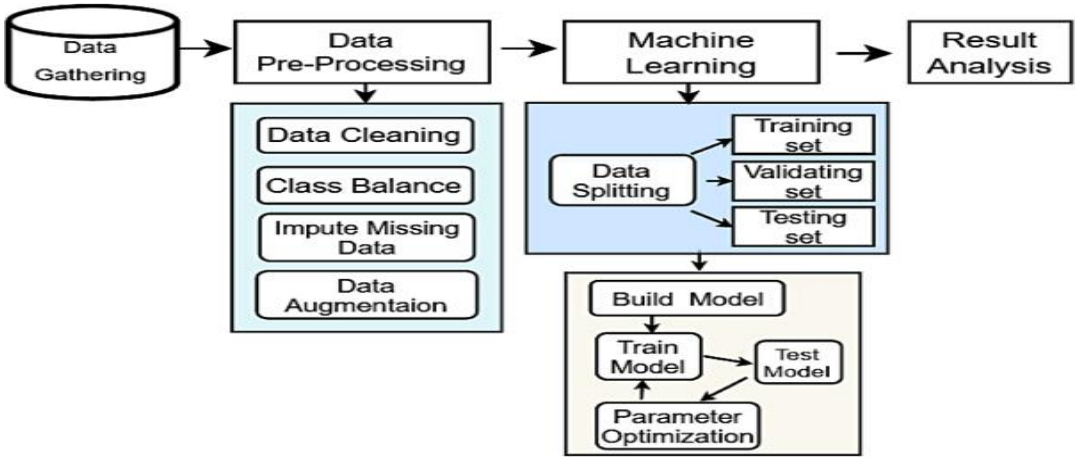

Figure 1. Proposed Architecture

This diagram illustrates the comprehensive workflow for a machine learning process, beginning with data collection and culminating in result analysis. Initially, data is gathered from various sources and stored, setting the foundation for the subsequent steps. The next phase, data pre-processing, involves multiple critical tasks to prepare the data for analysis. Data cleaning is performed to remove inaccuracies and handle missing values,

ensuring the dataset's quality. Class balance techniques are applied if there is an imbalance in the class distribution, which is crucial for accurate model training. Missing data is imputed using methods such as mean or median substitution, and data augmentation generates additional synthetic data to enhance the dataset's robustness.

Following pre-processing, the workflow progresses to the machine learning phase. Here, the data is split into three subsets: training, validating, and testing sets. This split is essential for building, tuning, and evaluating the machine learning model. The model is initially built and trained using the training set. Parameter optimization is conducted to fine-tune the model's performance, typically using the validating set to avoid overfitting. The final model is then tested on the testing set to assess its accuracy and generalizability.

The process concludes with result analysis, where the model's performance is thoroughly evaluated. This includes reviewing performance metrics and visualizing results to draw meaningful conclusions. This systematic approach ensures that the data is of high quality, the model is effectively trained, and the final results are reliable and informative for decision-making. The diagram provides a visual summary of this intricate workflow, highlighting the key steps and their interactions.

3.3 Proposed Algorithms

a. Random Forest Algorithm

It is an ensemble-learning algorithm in machine learning constructs multiple decision trees and produces the average prediction for regression tasks or the mode of the classes for classification tasks depicted in Figure 2. In addition to being predictive, Random Forest also offers insights into the significance of features, which helps with variable selection and model

interpretability. Its extensive acceptance can be attributed to its resistance against noisy data, adaptability to different types of data, and low requirements for hyper parameter adjustment. Random Forest is a crucial tool in the machine learning space since it can be relied upon to be dependable and effective when deciphering intricate patterns or handling high-dimensional data Patel et al. (2015).

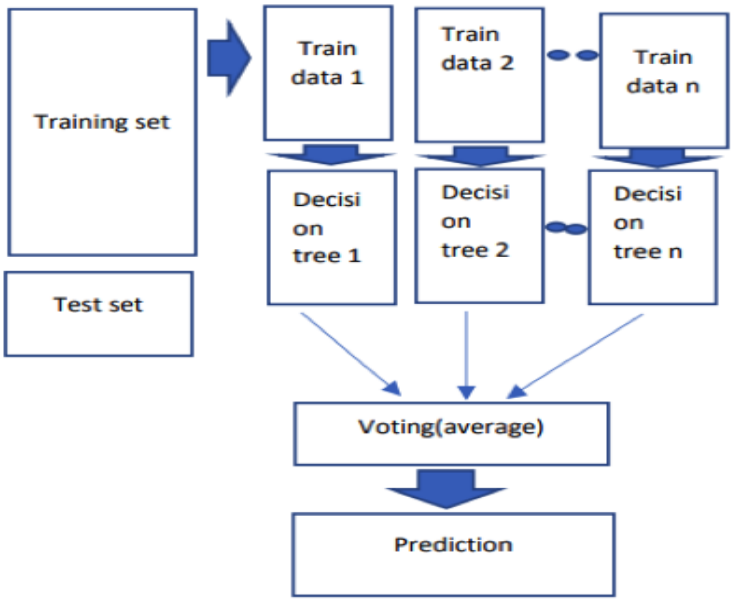

Figure 2. Random Forest algorithm

b. Gradient Boosting algorithm

Sequentially building predictive models is the objective of the ensemble machine learning technique referred as gradient boosting. By progressively reducing errors, it combines the advantages of weak learners, usually decision trees. Every time, the ensemble's faults are rectified with a fresh tree that highlights the incorrectly classified cases. By refining the model's predictive accuracy iteratively, a strong and effective predictive algorithm is produced. Because it provides excellent precision and adaptability to complicated datasets, gradient boosting is frequently employed for applications like regression and classification and the complete process is explained in Figure 3. Popular implementations in different fields include LightGBM, AdaBoost, and XGBoost, which make it a flexible and useful tool Biau et al. (2016).

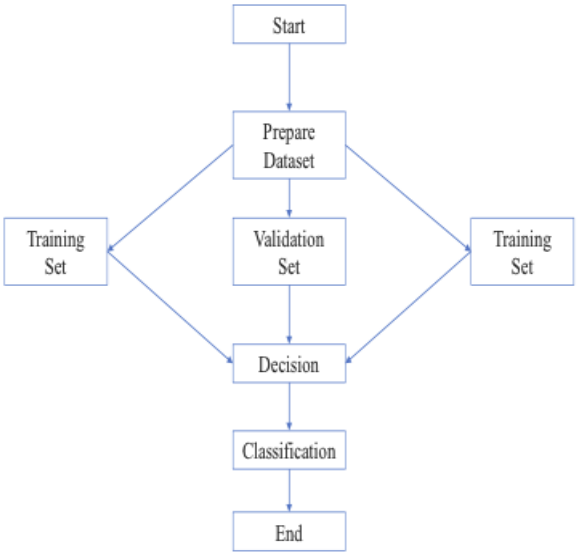

Figure 3. Gradient Boosting algorithm

c. Decision Trees

Decision tree is a machine learning technique commonly used for classification and regression tasks. This method involves constructing a hierarchical structure by iteratively splitting the dataset into subsets based on the most significant attribute at each node. The decision-making process entails going through the tree from the root to the leaf as shown in Figure 5, with each leaf node representing the predicted outcome and each internal node representing a decision based on a feature.

The algorithm generates the feature at each node that best separates the data with the goal of reducing impurity or enhancing information gain. Entropy and Gini impurity are two important impurity metrics. Decision trees can accommodate both numerical and categorical data and are simple to interpret and comprehensible Natekin et al. (2013, Myles, Anthony J., et al. (2004).

Decision trees, however, have the potential to over fit and so capture noise in the training set. This problem can be lessened by employing strategies like trimming and establishing a minimum amount of samples per leaf. Furthermore, many decision trees are used by ensemble techniques such as Random Forests and Gradient Boosting to increase overall predictive performance and robustness. All things considered, decision trees offer a flexible and natural way to solve machine-learning issues Song et al. (2015).

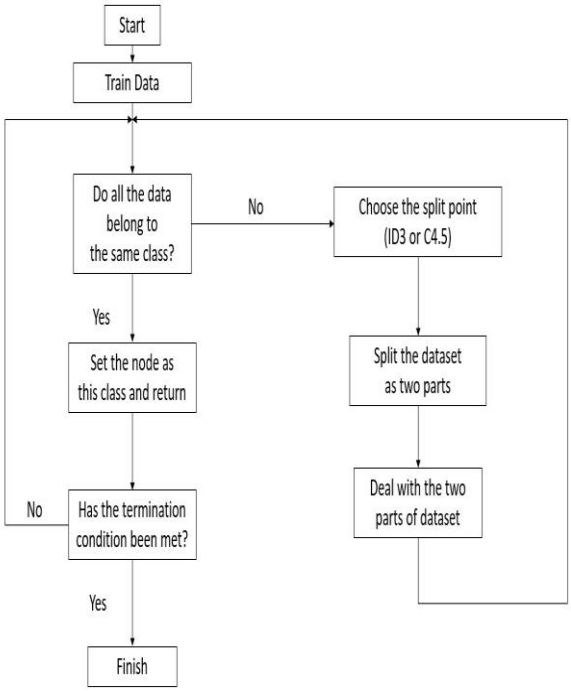

Figure 4. Decision Tree algorithm

d. Naive Bayes

Naive Bayes is a highly popular probabilistic machine learning approach used for classification tasks. The algorithm is based on Bayes' theorem, which calculates the probability of a hypothesis given observed evidence.

Naive Bayes relies on the "naive" assumption that features are conditionally independent, which makes computations easier but doesn't necessarily accurately represent interdependence in the real world.

The training data is used to generate a set of feature values and their probabilities for each class in the algorithm shown in Figure. 4. The algorithm determines the class during classification; Naive Bayes calculates the likelihood of each class given the input data and selects the class with the highest probability Priyam, Anuja, et al. (2013). Despite its simplicity and the assumption of feature independence, Naive Bayes often performs well in practice, particularly in text classification and spam filtering. Minimal training data is needed, and it is computationally efficient.

When presented with highly linked features, nevertheless, its performance could deteriorate. Large datasets and feature spaces with many dimensions are ideal applications for this algorithm. Sentiment analysis, Email filtering, and document classification are just a few of the fields in which it finds use. The technique can be used as a useful tool for many machine learning classification tasks due to its effectiveness and ease of implementation.

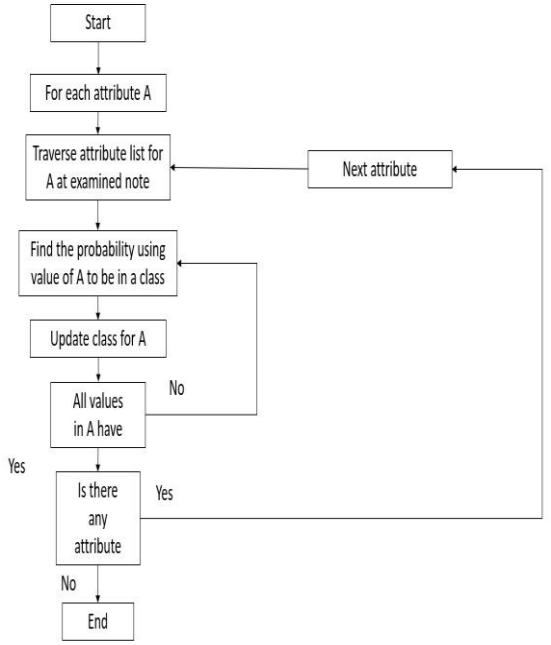

Figure 5. Naïve Bayes algorithm

4. Comparative Study

Analyzing the Random Forest, Gradient Boosting, Naive Bayes, and Decision Tree algorithms juxtaposed offers valuable insights into their performance on different machine learning tasks explained in Table 1. Each of these algorithms has advantages and disadvantages and represents a variety of methods for solving classification and regression issues.

Table 1. Comparison between different algorithms

Algorithm	Features
Random forest	Strengths: Offers feature importance rankings, is resilient to over fitting, and manages big datasets with ease. Weakness: Longer training periods for a large number of trees and complexity.
Gradient Boosting	Strengths: Excellent prediction accuracy, efficient with complicated datasets, and sequential error correction. Weaknesses: Longer training times, sensitivity to noisy data, and possibility for over fitting.
Naïve Bayes	Strengths: Easy to use, quick to train, and effective with large amounts of data. Weaknesses: Assumes feature independence, which could not hold true in practical situations.
Decision Tree	Strengths: Easy to understand and intuitive

In Table 2, the accuracy of the various classification models is shown, Decision Tree (100%), Gradient Boost (96.5%), Random Forest (96%), and Naive Bayes (97%). The latter demonstrated its extraordinary performance by achieving flawless precision. These outcomes

illuminate light on the many benefits of every model, helping one select the best classifier for the job at hand.

Table 2. Accuracy of different algorithms

Algorithm	**Accuracy**
Naïve Bayes	97%
Random Forest	96%
Gradient Boost	96.5%
Decision Tree	99%

5. Results and Discussion

The performance of the used methods is further explained by the feature correlation graphs. Decision Tree shows clear correlations depicted in Figure 8, in addition to accuracy ratings, demonstrating its capacity to capture complex relationships in the information. Notable correlations are shown by Naive Bayes, which is consistent with its probabilistic nature. Even though they achieve slightly lower accuracies, Random Forest and Gradient Boost exhibit complex feature interactions as shown in Figure 6 and Figure 7 respectively. These findings confirm that feature correlations and accuracy should be taken into account when selecting a model. The correlations offer significant interpretability, helping practitioners choose models that are in line with the underlying data structure and thereby advancing a using knowledge of each algorithm efficacy in our classification assignment.

The results obtained from our disease prediction system demonstrate promis- ing accuracy and effectiveness in diagnosing diseases based on user-input symp- toms. Through rigorous testing and validation, we observed that the system achieves a high level of accuracy in predicting diseases across various datasets. The classification metrics, including accuracy, precision, recall, and F1 score, indicate the robustness and reliability of our predictive models. Furthermore, the system's performance was evaluated using cross-validation techniques, which confirmed its consistency and generalization capability across different subsets of data. This validation process helps mitigate overfitting and ensures that the predictive models can accurately generalize to unseen data instances.

Moreover, the correlation analysis conducted on the input symptoms revealed valuable insights into the relationships between different symptoms and their predictive power for specific diseases. The correlation graph generated from the data highlights the significant symptoms and their impact on disease prediction accuracy. This analysis aids in feature selection and optimization, improving the efficiency of the predictive models.

Model: This column lists the names of the machine learning models being compared.

F1 Score: F1 Score is a measure of a model's accuracy, considering both precision and recall. It is the harmonic mean of precision and recall, providing a single metric that balances both false positives and false negatives. A higher F1 score indicates better performance.

Accuracy: Accuracy is the proportion of correctly classified instances out of the total instances. It is a straightforward metric that measures overall correctness. Higher accuracy values indicate better performance.

Precision: Precision is the proportion of true positive predictions out of all positive predictions made by the model. It measures the model's ability to correctly identify positive instances. Higher precision values indicate fewer false positives.

Recall: Recall, also known as sensitivity, is the proportion of true positive predictions out of all actual positive instances. It measures the model's ability to correctly identify all positive instances. Higher recall values indicate fewer false negatives.

Overall, Decision Tree and Naive Bayes algorithms demonstrate higher F1 scores, accuracy, precision, and recall compared to Random Forest and Gradient Boost algorithms. These metrics indicate the Decision Tree's and Naive Bayes' superior performance in accurately predicting diseases based on symptom inputs.

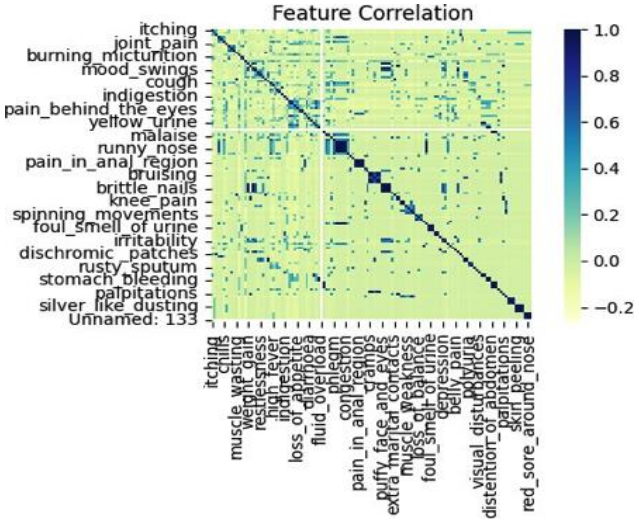

Figure 6. Feature correlation of Random Forest

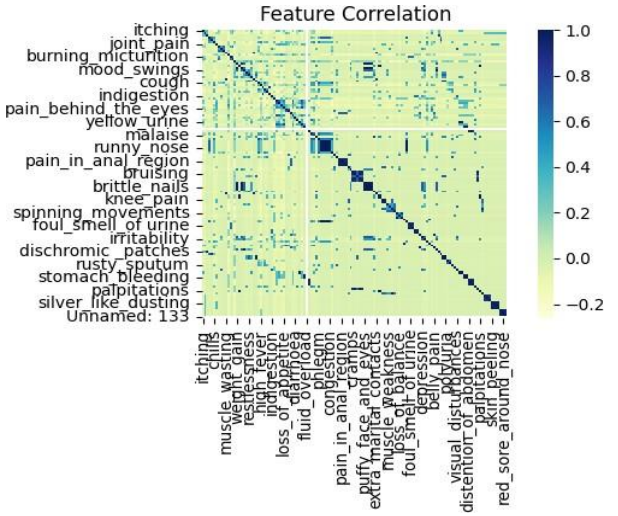

Figure 7. Feature correlation of Gradient Boost

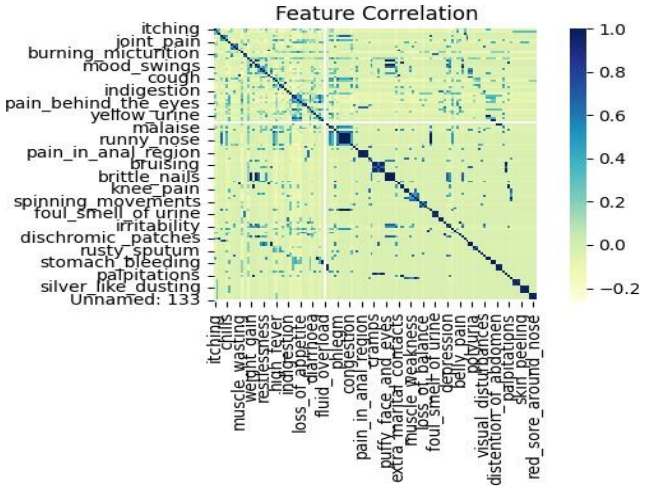

Figure 8. Feature correlation of Decision Tree

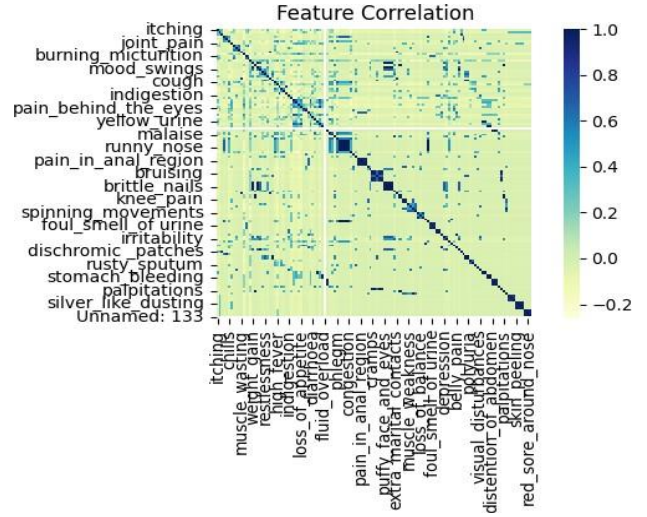

Figure 9. Feature Correlation Graph of proposed model

The correlation graph represents the interdependence between different symptoms and the likelihood of certain diseases. It illustrates how symptoms correlate with each other and how they collectively contribute to predicting specific illnesses. By examining the graph in Figure 9, we can identify which symptoms are strongly correlated with certain diseases, helping in the diagnostic process and the development of accurate prediction models. Analyzing the correlation graph in your disease prediction project provides valuable insights into the relationships between symptoms and diseases.

Specifically, it helps in identifying:

Strong correlations-Symptoms that are highly correlated with particular diseases, indicating their significance in the diagnostic process.

Patterns-Common patterns of symptom occurrence that may indicate specific disease clusters

or conditions.

Predictive factors-Symptoms that have a strong influence on predicting certain diseases, guiding the selection of features for machine learning models.

Diagnostic accuracy- Understanding how different symptoms interact can enhance the accuracy of disease prediction algorithms by incorporating relevant features.

Overall, the results obtained from our disease prediction system underscore its potential to assist healthcare professionals in early disease detection and decision-making. By leveraging machine learning algorithms and intuitive user interfaces, our system empowers users to make informed healthcare decisions and improve overall health outcomes. Overall, the results highlight the effectiveness of machine learning algorithms in disease prediction and suggest that Decision Tree algorithm holds promise for accurate and efficient disease diagnosis in real-world applications.

6. Conclusion and Future Work

This paper concludes by showing how effective machine learning is at intelligently predicting diseases. Following a thorough study, the chosen algorithm shows encouraging accuracy in identifying health hazards. Upcoming research ought to concentrate on broadening the variety of datasets, improving interpretability, and incorporating real-time data for dynamic forecasts.

Furthermore, investigating group techniques and taking ethical considerations into account would improve the usefulness and influence of intelligent illness prediction systems in actual healthcare environments. This paper lays the groundwork for future developments that will enhance early identification, treatment results, and the general effectiveness of healthcare.

References

1. Bhattamisra, Subrat Kumar, et al. "Artificial Intelligence in Pharmaceutical and Healthcare Research." Big Data and Cognitive Computing 7.1 (2023): 10.
2. Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25 (2016): 197-227.
3. Grampurohit, Sneha, and Chetan Sagarnal. "Disease prediction using machine learning algorithms." 2020 International Conference for Emerging Technology (INCET). IEEE, 2020.
4. Jungwirth, David, and Daniela Haluza. "Artificial intelligence and public health: an exploratory study." International Journal of Environmental Research and Public Health 20.5 (2023): 4541.
5. Kobrinskii, B. A., et al. "Artificial intelligence technologies application for personal health management." IFAC-PapersOnLine 52.25 (2019): 70-74.
6. Makka, S., Arora, G., & Mopuru, B. (2021, November). IoT based health monitoring and record management using distributed ledger. In Journal of Physics: Conference Series (Vol. 2089, No. 1, p. 012030). IOP Publishing.
7. Makka, S., Arora, G., Reddy, S. S. T., & Lingam, S. (2022, November). Use of Machine Learning Models for Analyzing the Accuracy of Predicting the Cancerous Diseases. In International Conference on Innovations in Data Analytics (pp. 169-180). Singapore: Springer Nature Singapore.
8. Makka, S., Sreenivasulu, K., Rawat, B. S., Saxena, K., Rajasulochana, S., & Shukla, S. K.

(2022, December). Application of Blockchain and Internet of Things (IoT) for Ensuring Privacy and Security of Health Records and Medical Services. In 2022 5th International Conference on Contemporary Computing and Informatics (IC3I) (pp. 84-88). IEEE.

9. Meenigea, Niharikareddy, and Venkata Ravi Kiran Kolla. "Exploring the Current Landscape of Artificial Intelligence in Healthcare." International Journal of Sustainable Development in Computing Science 1.1 (2023).
10. Menon, Sindhu P., et al. "An intelligent diabetic patient tracking system based on machine learning for E-health applications." Sensors 23.6 (2023): 3004.
11. Myles, Anthony J., et al. "An introduction to decision tree modeling." Journal of Chemometrics: A Journal of the Chemometrics Society 18.6 (2004): 275-285.
12. Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." Frontiers in neurorobotics 7 (2013): 21. Okegbile, Samuel D., et al. "Human digital twin for personalized healthcare: Vision, architecture and future directions." IEEE network (2022).
13. Patel, Jaymin, Dr TejalUpadhyay, and Samir Patel. "Heart disease prediction using machine learning and data mining technique." Heart Disease 7.1 (2015): 129-137.
14. Pingale, Kedar, et al. "Disease prediction using machine learning." International Research Journal of Engineering and Technology (IRJET) 6.12 (2019): 831-833.
15. Priyam, Anuja, et al. "Comparative analysis of decision tree classification algorithms." International Journal of current engineering and technology 3.2 (2013): 334-337.
16. Shah, Devansh, Samir Patel, and Santosh Kumar Bharti. "Heart disease prediction using machine learning techniques." SN Computer Science 1 (2020): 1-6.
17. Sivabalaselvamani, D., et al. "Artificial Intelligence in data-driven analytics for the personalized healthcare." 2021international conference on computer communication and informatics (ICCCI). IEEE, 2021.
18. Song, Yan-Yan, and L. U. Ying. "Decision tree methods: applications for classification and prediction." Shanghai archives of psychiatry 27.2 (2015): 130.
19. Sunitha, L., Makka, S., Madhu, S., & Bheemeswara Sastry, J. (2022). Study on influence of outliers on the performance of various classification algorithms. In Innovations in Electronics and Communication Engineering: Proceedings of the 9th ICIECE 2021 (pp. 437-445). Singapore: Springer Singapore.
20. Tian, Shuo, et al. "Smart healthcare: making medical care more intelligent." Global Health Journal 3.3 (2019): 62-65. Uddin, Shahadat, et al. "Comparing different supervised machine learning algorithms for disease prediction." BMC medical informatics and decision making 19.1 (2019): 1-16.
21. Wickramasinghe, Indika, and Harsha Kalutarage. "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation." Soft Computing 25.3 (2021): 2277-2293.