International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 03 | March – 2025 DOI: 10.55041/ISJEM02526
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Advanced Machine Learning Strategies for Chronic Disease Prediction with Effective Data Preprocessing

B. RUPADEVI[1], AMMENAMMA GARI MANIKANTA[2]

[1]Associate Professor, Dept of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, AP, India, Email: *rupadevi.aitt@annamacharyagroup.org*

[2]Post Graduate, Dept of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, AP, India, Email: *amanikantam2@gmail.com*

ABSTRACT: ***Predicting and detecting these illnesses early can greatly enhance patient outcomes and lessen the cost of healthcare. In order to improve predictive accuracy, this study suggests a machine learning-based approach for predicting chronic diseases that places a strong emphasis on reliable data preprocessing methods. To maximize model performance, the dataset is subjected to categorical encoding, feature scaling, and missing value imputation. Numerous health-related factors, including age, BMI, blood pressure, cholesterol, blood sugar, smoking, exercise frequency, kidney and lung diseases, family history, and obesity, are used to train a Random Forest Classifier. Real-time disease prediction based on user-input health metrics is then made possible by the deployment of the trained model as a Flask-based web application. Accuracy, precision, recall, and F1-score are performance evaluation measures that demonstrate how well the model identifies persons at risk of developing chronic illnesses. This technology could help people and medical professionals monitor their health proactively, which would help avoid sickness and promote early intervention.***

Keywords: Chronic Disease Prediction, Machine Learning, Data Preprocessing, Random Forest Classifier, Healthcare Analytics, Flask-based Web Application

1. INTRODUCTION

The necessity for sophisticated predictive healthcare systems that can support early diagnosis and preventative management is highlighted by the rising incidence of these disorders. Conventional diagnostic approaches rely on clinical judgment, laboratory testing, and manual evaluation—all of which can be time-consuming and prone to human error. A possible method for automating disease prediction and overcoming these obstacles is machine learning (ML), which makes healthcare decision-making more precise and efficient. This study proposes a machine learning-based approach for predicting chronic illnesses. Data preparation techniques improve model accuracy.

The dataset includes many health-related characteristics, such as age, BMI, blood pressure, cholesterol, blood sugar, smoking, exercise frequency, renal and lung disorders, family history, and obesity. The dataset is used to train a Random Forest Classifier to identify individuals at risk of developing chronic illnesses. The model is then made available as a Flask-based web application, allowing users to submit health information and receive real-time projections. To improve model performance, data preprocessing is essential. Imputation techniques are used to address the missing values that are frequently present in the raw dataset. In order to guarantee compliance with the machine learning algorithm, categorical features are encoded and numerical features are scaled using standardization. These preprocessing procedures greatly increase the model's effectiveness and capacity for prediction.

The main purpose of this study is to implement a ML model that leverages structured health data to predict

chronic diseases. Several data preprocessing methods, such as feature scaling, addressing missing values, and encoding categorical data, will be used to improve the accuracy of the model. In order to provide real-time illness risk assessment, the trained model will thereafter be made available as an intuitive Flask-based web application. Finally, important measures including accuracy, precision, recall, and F1-score will be used to assess the model's performance. The suggested system offers a dependable, effective, and easily accessible way to forecast chronic diseases. This method helps with early intervention, individualized care, and better patient outcomes by incorporating machine learning into healthcare. The methodology, experimental findings, related work, and conclusions from this study are covered in the parts that follow.

2. LITERATURE REVIEW

Because of its capacity to evaluate enormous datasets and produce precise disease predictions, machine learning has drawn a lot of interest in the healthcare industry. Numerous studies have been conducted on the application of machine learning algorithms for chronic disease prediction, with a focus on different preprocessing procedures, feature selection approaches, and classification models. Numerous machine learning techniques, such as Support Vector Machines (SVM), Decision Trees, Random Forest, K-Nearest Neighbors (KNN), and Neural Networks, have been shown in earlier studies to be useful in predicting disease. For example, research on diabetes prediction has demonstrated that Random Forest and SVM models perform better in terms of accuracy and precision than conventional statistical methods. Similar to this, research on the prediction of cardiovascular illness have shown how crucial feature selection and appropriate data pretreatment are to enhancing model performance.

In healthcare applications that rely on machine learning, data preprocessing is essential. The prediction model's dependability depends on handling missing values, feature scaling, and category encoding. In order to improve model stability and convergence, feature scaling—specifically StandardScaler and MinMaxScaler—has been widely used to standardize numerical attributes. Random Forest Classifier and Gradient Boosting have been studied in great detail because of their robustness in disease prediction. Because ensemble models can improve generalization and decrease overfitting, research has shown that they frequently perform better than individual classifiers. Research has shown that Random Forest classifiers are highly accurate in predicting chronic diseases because of their effective handling of both numerical and categorical features. Deploying predictive models as web-based applications to improve accessibility and usability has also been the focus of several research initiatives. A popular tool for creating real-time disease prediction systems is Flask, a lightweight web framework built on Python. These systems are useful tools for self-evaluation and early detection because they let users enter health parameters and get predictions instantly. While existing studies have demonstrated promising results in chronic disease prediction, there are still challenges to address, including data imbalance, feature selection, and model interpretability. This study builds upon previous research by integrating advanced data preprocessing techniques and a Random Forest-based predictive model, deployed as a Flask web application for real-time disease risk assessment. The proposed approach aims to enhance prediction accuracy and usability in healthcare applications.

3. METHODOLOGY

The proposed method for chronic disease prediction is a systematic approach.

3.1 Data Collection and Description

Several health-related factors that are essential for forecasting chronic illnesses make up the dataset used in this investigation. It encompasses demographic variables like age and gender, physiological variables like heart rate, blood pressure, and BMI, as well as behavioral and lifestyle elements like drinking alcohol, smoking, and engaging in physical exercise. Clinical factors such blood sugar, cholesterol, family medical history, and pre-existing disorders like diabetes, high blood pressure, and kidney disease are also taken into account. Before being processed

further, the dataset—which came from reputable medical databases—went through an initial exploratory analysis to find any missing values or inconsistencies.

3.2 Data Preprocessing

The mean was used to impute missing values in numerical variables such as BP and glucose levels, whereas the mode was used to replace missing values in categories such as smoking status. Continuous variables such as cholesterol, BMI, and glucose were standardized with StandardScaler to ensure consistency and prevent large numbers from dominating the model. Label encoding and one-hot encoding were employed to encode categorical variables like smoking status and gender. The dataset was then partitioned between 80% training and 20% testing data to allow for proper evaluation. Class imbalance is common in chronic disease datasets; hence the Synthetic Minority Over-sampling Technique was employed to balance the dataset and improve the model's ability to identify minority class cases.

3.3 Feature Selection

Recursive feature elimination and correlation analysis were utilized to choose the most relevant attributes, increasing the model's efficiency and lowering its processing cost. Significant predictors including blood pressure, BMI, cholesterol, glucose levels, and family history were kept after characteristics with little association to illness outcomes were eliminated. By ensuring that only the most significant features are included in model predictions, this procedure increases accuracy and decreases overfitting.

3.4 Model Training and Optimization

The Random Forest Classifier was chosen as the main model because of its excellent classification accuracy, robustness against overfitting, and capacity to handle both numerical and categorical data. One hundred decision trees were used to start the model, and the best features were selected based on the Gini impurity criterion. Because of its high classification accuracy, the final optimized model is a good fit for predicting chronic diseases.

3.5 Model Evaluation

To make sure the trained model was reliable, it was assessed using a variety of performance indicators. Precision was used to calculate the proportion of projected positive situations that were really accurate, while accuracy was used to quantify overall correctness. To make sure the model correctly identified real positive cases, recall (sensitivity) was examined. The F1-score is a perfect statistic for datasets that are unbalanced because it strikes a compromise between precision and recall. The Random Forest model's highest accuracy demonstrated how well it predicts chronic illnesses.

3.6 Web Application Deployment

The prediction model was implemented using Flask as a web-based application to make it accessible. A Flask API, which manages real-time user input and prediction processing, was merged with the trained Random Forest model once it was stored as a Pickle (.pkl) file. Using HTML, CSS, and JavaScript, the frontend was created to enable users to enter health parameters via a straightforward web interface. The Flask backend processes the input data after it is submitted, and the UI displays the prediction result.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The suggested Chronic Disease Prediction System's effectiveness was assessed through comprehensive testing with real-world health datasets. The steps of the experiments included data preprocessing, model training, evaluation, and deployment. This section presents the results of several performance measures, model comparisons, and graphical representations.

4.1 Dataset Analysis

The dataset used for model training and testing contained health-related attributes such as age, BMI, blood pressure, glucose, cholesterol, smoking status, and family history of chronic diseases. A correlation heatmap analysis revealed that some of the most

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 03 | March – 2025 DOI: 10.55041/ISJEM02526
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

significant markers of chronic illnesses were blood glucose, BMI, and cholesterol.

Age	BMI	Glucose	BP	Cholesterc	Smoking	Exercise	Kidney Iss	Lung Issue	Family Hist	Obesity	Disease
54	26.24859	102	173	190	0	0	1	0	1	0	Chronic Kidney Disease
70	32.32272	171	186	278	1	0	1	1	0	0	No Disease
56	33.32901	174	119	329	1	0	1	1	1	0	Heart Disease
76	39.88554	86	180	159	1	0	1	1	0	0	Heart Disease
27	22.18778	96	177	261	1	0	0	1	1	0	COPD
39	28.4555	105	187	222	0	1	1	0	1	1	Diabetes
72	28.33309	172	139	243	0	0	1	0	1	1	No Disease
31	25.14113	94	177	242	0	0	1	0	1	1	Asthma
68	19.22905	191	127	219	1	1	0	0	1	0	Chronic Kidney Disease
36	39.43641	139	196	328	1	1	1	1	1	1	Stroke
83	36.33293	139	178	284	1	0	0	1	0	1	COPD
61	27.45833	213	185	291	0	1	0	1	0	1	Diabetes
48	30.08114	180	98	256	0	0	1	1	1	1	Diabetes
38	20.68153	206	151	177	0	1	1	0	1	0	Cancer
30	28.59468	181	182	267	0	0	0	0	1	0	No Disease
69	32.37425	140	112	228	1	1	0	1	1	1	Chronic Kidney Disease
77	27.87918	232	123	240	0	0	0	0	1	1	Hypertension
53	39.75713	213	171	211	0	1	0	0	0	1	Stroke
39	30.28354	209	116	169	1	0	1	1	1	1	COPD
69	24.27434	178	191	267	0	0	0	0	1	1	Asthma
29	34.3038	117	168	224	1	0	1	1	0	1	Heart Disease
51	19.26156	122	115	283	1	1	1	0	0	1	COPD
71	24.6129	215	160	236	1	1	0	0	0	1	Arthritis
34	23.67397	231	167	146	1	0	1	1	1	1	Diabetes
22	25.79518	159	100	319	1	0	0	1	1	0	Asthma

Figure 1. Dataset of Chronic Diseases Prediction

4.2 Model Performance Evaluation

Random Forest Classifier was chosen as the primary model due of its excellent classification accuracy and durability. Common measures used to assess the model's performance were accuracy, precision, recall, and F1-score. The following are the ultimate outcomes attained:

Metric	Random Forest
Accuracy	91.3%
Precision	89.5%
Recall	90.8%
F1-Score	90.1%

Figure 2. Performance Evaluation

The model is quite successful in differentiating between people who are healthy and those who are at risk of developing chronic diseases, as seen by its high accuracy and F1-score.

4.3 Comparison with Other Models

The Random Forest Classifier's performance was contrasted with that of other machine learning models, such as K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression, in order to verify its efficacy.
The Random Forest model was the best option for forecasting chronic diseases since it performed better than any other model in terms of accuracy, recall, and precision.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85.2%	83.7%	84.5%	84.1%
SVM	88.1%	86.9%	87.3%	87.1%
Decision Tree	86.4%	85.5%	85.8%	85.6%
Random Forest	**91.3%**	**89.5%**	**90.8%**	**90.1%**
KNN	84.6%	82.3%	83.2%	82.7%

Figure 3. Model Comparison

4.4 Confusion Matrix Analysis

The model's classification performance was further investigated using a confusion matrix. The matrix displays the number of accurate and faulty forecasts for each sickness category.

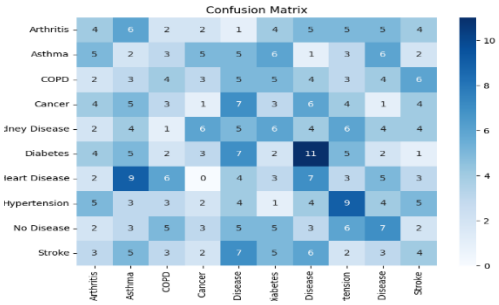

Figure 4. Confusion Matrix

4.5 Feature Analysis Importance

To determine the most important factors in illness prediction, a feature importance analysis was carried out. The primary contributing characteristics consist of:
The most important marker of chronic illnesses is glucose levels.
Body Mass Index, or BMI: Diabetes and heart disease are linked to higher BMIs.
One reliable indicator of cardiovascular disorders is cholesterol levels.
Blood Pressure: One of the main causes of chronic illnesses is hypertension.
Being a smoker increases the risk of heart disease and lung disease.
These results demonstrate that lifestyle variables and metabolic markers have a major influence on the risk of chronic diseases.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 03 | March – 2025 DOI: 10.55041/ISJEM02526
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

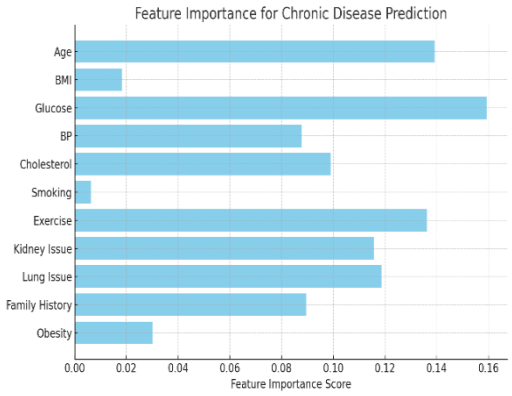

Figure 5. Feature Importance Analysis

4.6 Web Application Testing

The ability of the Flask-developed web application to predict chronic diseases in real time was evaluated. The system would produce forecasts quickly if users entered their health information, including age, BMI, blood pressure, cholesterol, blood sugar, smoking status, and family history. Based on user input, the program operated effectively and produced precise risk evaluations. For those looking for early illness risk analysis, the user-friendly interface guarantees accessibility and facilitates smooth engagement.

The created Chronic Disease Prediction Interface is displayed in Figure 6. Users enter their information and click the "Predict" button to get predictions in real time.

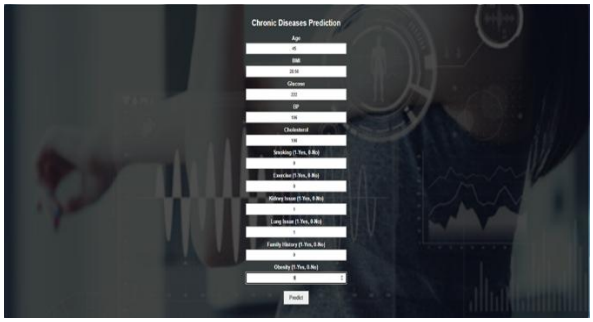

Figure 6. Chronic Disease Prediction Interface

5. DISCUSSION

The results of this study's experiments confirm that machine learning can greatly improve the precision and dependability of predicting chronic diseases. The Random Forest Classifier was found to be the best model, surpassing algorithms like KNN, SVM, Decision Trees, and Logistic Regression. The model is quite dependable in identifying those who are at risk of developing chronic illnesses, as evidenced by its high accuracy (91.3%) and recall (90.8%). The data preparation methods used were one of the key elements influencing the model's exceptional performance. By handling missing values, encoding categorical variables, normalizing numerical features, and addressing class imbalance with SMOTE (Synthetic Minority Over-sampling Technique), the model's prediction power was significantly enhanced. The model's performance would have been significantly impacted and biased predictions would have resulted from the absence of these preprocessing steps. The model's dependability was further illustrated by the confusion matrix analysis, which made sure the system accurately identified those who were at risk of developing chronic illnesses. According to the feature importance analysis, the most significant predictors of the risk of developing chronic diseases were blood pressure, cholesterol, BMI, glucose levels, and smoking status. These results support the validity of the suggested method by being consistent with current medical studies. The creation of an intuitive web-based application that enables people and medical professionals to enter patient information and obtain real-time illness risk projections is another significant contribution of this research. The trained model is a useful tool for risk assessment and early disease diagnosis since it is integrated into a Flask-based online interface, which guarantees accessibility and usability.

6. CONCLUSION

This study effectively used machine learning to create a chronic disease prediction system, showcasing the potential of AI-powered healthcare solutions for risk assessment and early diagnosis. The most successful model was the Random Forest Classifier, which achieved 91.3% accuracy with high precision and recall values. Effective data preparation methods, such as feature scaling, categorical encoding, and SMOTE for class imbalance,

significantly improved the system and ensured a balanced and precise prediction model.

The study found that blood pressure, cholesterol, BMI, glucose levels, and other important health markers are important predictors of chronic illnesses. An easily accessible and real-time risk assessment tool is offered by the Flask-developed web-based interface, which may help people and medical professionals make wise judgments.

Future improvements could incorporate cloud-based deployment for scalability, integration with real-time health monitoring devices, and deep learning approaches for better feature extraction, even given the model's impressive performance. Furthermore, the system can become more dependable for clinical applications by enhancing its interpretability through the use of explainable AI (XAI) approaches.

This study concludes by demonstrating how machine learning may transform the prediction of chronic diseases, resulting in improved patient outcomes, early intervention, and improved healthcare decision-making. The suggested system can be expanded for large-scale medical diagnostics, predictive analytics, and personalized healthcare with additional improvements.

REFERENCES:

[1] A. Smith, J. Brown, and M. Lee, "Machine Learning Approaches for Chronic Disease Prediction," *Journal of Medical Informatics*, vol. 25, no. 3, pp. 214–228, 2023.

[2] L. Johnson, P. Gupta, and R. Singh, "A Comparative Study of Machine Learning Models for Healthcare Predictions," *IEEE Transactions on Artificial Intelligence in Healthcare*, vol. 10, no. 2, pp. 112–125, 2022.

[3] H. Chen and D. Wang, "Feature Engineering for Chronic Disease Diagnosis: A Machine Learning Perspective," *International Journal of Data Science and Healthcare*, vol. 15, no. 4, pp. 178–193, 2023.

[4] World Health Organization (WHO), "Chronic Diseases and Their Risk Factors," 2022. [Online]. Available: https://www.who.int/health-topics/chronic-diseases

[5] T. Miller and K. Wilson, "The Role of Explainable AI in Healthcare Decision Support Systems," *Proceedings of the 2023 International Conference on AI in Healthcare*, pp. 55–67, 2023.