European Economic Letters
ISSN 2323-5233
Vol 15, Issue 4 (2025)
http://eelet.org.uk

Uncovering Diagnostic Patterns: Exploratory Machine Learning Approaches for Medical Condition Classification

Paromita Sadhu*, Arpit Rastogi *, Ayush Sinha*, Darshini R*,
2nd Year PGDM Student, Xavier Institute of Management & Entrepreneurship, Chennai

ABSTRACT

Machine learning has huge prospects in disease early detection and enhanced treatment of patients. The present paper presents a predictive model of medical states by use of a Kaggle healthcare dataset of demographic, clinical, and lifestyle variables. The most significant health indicators will be age, glucose level, blood pressure, BMI, oxygen saturation, cholesterol, triglycerides, HbA1c, and behavioral indicators, such as smoking, alcohol use, and physical exercises. Four classification models were taken into account to identify the most effective classification model; these were Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree and Random Forest. The model performance metrics were accuracy, F1-Score and ROC-AUC. The precision obtained with the Logistic Regression was 91.37%, Decision Tree 84.16%, Random Forest 91.63% and KNN 81.53%. ROC-AUC results indicated a high predictive value among all the models and in most circumstances of significance An AUC value above 0.95 in Logistic Regression and KNN. Random Forest also gave values of AUC exceeding 0.95 in all disorders of significance and more challenging to identify multi-class disorders, such as cancer, diabetes and asthma. The findings indicate that the ensemble-based methods can be viable when compared to the traditional classifiers, when dealing with healthcare data, which is both non-linear and high-variance. Overall, the proposed predictive models can have great potential when used as a source of clinical decisions and preventive healthcare.

Keywords— Machine Learning, Predictive Modeling, Medical Diagnosis, Healthcare Analytics, Random Forest, Logistic Regression

1. INTRODUCTION

AI and ML are revolutionizing modern healthcare through the proficiency of intelligent systems to interpret complex medical data and support clinicians in decision-making. These are facilities that open doors to predictive analytics, enabling healthcare organizations to transition from a predominant disease management approach to one of prevention and early intervention. These are facilities that open doors to predictive analytics, enabling healthcare organizations to transition from a predominant disease management approach to prevention and early intervention. Given the widespread adoption of electronic health records, wearable sensors and diagnostic data which are being produced at an ever-increasing rate, ML now plays a critical role in finding insight that lies hidden within, predicting risks for diseases and guiding patient-specific health interventions. The dataset that was used for this study was obtained from Kaggle, and included several demographic, physiological and lifestyle variables: age, gender, glucose level, blood pressure, body mass index (BMI), oxygen saturation, cholesterol, triglycerides, HbA1c, stress level, physical activity, and diet score [1].

This set of features reflects the multidimensionality of human health, where various medical, behavioral and genetic factors jointly contribute to overall well-being. In addition, most classical diagnostic methods are dependent on clinical opinions and do not provide a good estimate of the nonlinear relationships between multiple parameters. However, it is still one of the most challenging problems in health analytics that how accurate and timely prediction can be done on individual based measurable indicators. This paper is an attempt toward the AI-based prediction system to predict a person's disease status, based on his physiological and lifestyle features, which are associated with various diseases like diabetes, hypertension, obesity and asthma. In this paper, a data-driven approach is adopted to build an AI-based predictive model that can improve the prediction accuracy on clinical conditions and thus would enable better-informed clinical decision making. Results in this work show that the application of AI/ML could potentially lead to significantly better healthcare management, earlier diagnosis and construction of real time health monitoring devices or systems with high reliability for early detection and prediction of medical risks.

2. LITERATURE REVIEW

The current level of scientific development in healthcare has unveiled a new world of data-driven and AI-powered healthcare to the patient. An example is cancer which is one of significant contributors of the worldwide burden of mortality and morbidity, and individual treatment is still an issue that is yet to be resolved. Zhang et al. (2023) [2] emphasize the urgency of the Artificial Intelligence (AI) and Machine Learning (ML) remaking the cancer diagnosis, prognosis, and treatment choice. ML algorithms have demonstrated they can accurately predict a number of cancers including breast, brain, lung, liver and prostate cancer using large datasets and have demonstrated the ability to outperform traditional clinical practices in their ability to predict these cancers, thereby supporting early detection and personalized treatment choices.

The IOT devices, healthcare systems, mobile applications and urban infrastructure have generated large heterogeneous data with the Fourth Industrial Revolution. According to Sarker et al. (2021) [3], advanced analytics and ML can allow making data-driven decisions that are intelligent and possible by transforming such datasets into actionable information. Their discussion shows that data science has a very broad-based scope in various fields such as business, healthcare, cyber security and urban analytics and issues such as data integration, scalability and ethical clearance.

In the meantime, AI-driven healthcare systems are increasingly becoming dependent on analytics and mobile computing in order to enhance patient care. According to Sitaraman et al. (2025) [4], the availability of these technologies would contribute to real-time health monitoring, predictive disease modeling and personalized treatment when integrated. Although these systems for access and early warning have massive latent potentials of addressing new opportunities of access and early warning challenges, issues of data privacy, interoperability, and infrastructure also present problems that reiterate the urgency of allowing frameworks to make their potential of use a reality.

Combination of ML with technologies of IoT has also advanced health care applications especially in continuous monitoring and predictive analytics. Bharadwaj et al. (2021) [5] established that sensor data produced by different IoTs captures myriad health related even with increasing number of such devices recruited, the more data is recruited to be further processed by ML techniques that facilitate pattern detection and prediction of disease besides early warning systems of chronic diseases. They can also enhance the effectiveness of functioning and patient outcomes because dealing with such concerns as data privacy and computers complexity.

Combined, these works point out how AI, ML and IoT integration is transforming healthcare by facilitating predictive and personalized and intelligent healthcare in various health fields. Privacy, interoperability, model interpretability and robust infrastructure to support widespread adoption are also mentioned among the most frequent obstacles (Khan et al., 2018). The above literature review suggests that epidemiology of AI, machine learning and IoT is pushing the contemporary healthcare facilities to the predictions analytical, real-time monitoring and treatment customization [6]. Those technologies were already demonstrated to be successful in other medical spheres, oncology and cardiovascular up to chronic disease management and smart health system. Even the significant progress in the field of enhancing the accuracy of diagnosis, efficiency and patient outcomes, the issues of data protection, the privacy concerns, system integration and algorithm bias remain as yet a challenge [7]. Further research in the future must focus on the improvement of explainable, interoperable and scalable AI solutions, which are clinically adoptable and morally acceptable. With such obstacles resolved, healthcare will be able to realize the full potential of data-driven technologies to deliver proactive, accurate and patient-centered care.

3. METHODOLOGY

The research method is a systematic framework to forecast the medical status of an individual using AI and machine learning algorithms. It consists of steps such as data collection, pre-processing, feature extraction/selection, model construction and assessment. The approach ensures that the dataset is preprocessed to be a clean, balanced, and analyzable format for predictive modeling in healthcare analytics.

3.1 Data Collection

The dataset used in this research was from Kaggle, uploaded by Abdallu H Ahmed [1]. This is a synthetic dataset created for educational and research purposes. It is not real patient records and has no personally identifiable information (PII). A random value mapping that corresponds to the ranges observed in clinical settings (Glucose, Blood pressure, Cholesterol and Body mass index) have been used for generating this simulated dataset. Some of the columns, like random notes and noise col were deliberately added as a noise to allow users to practice data cleaning and preprocessing. The dataset contains numerous clinical and controlling factors associated with human health. Each record represents an individual's medical and lifestyle profile, contains several characteristics like Age, Gender, Glucose, Blood Pressure, BMI, Oxygen Saturation, Cholesterol, Triglycerides, HbA1c, Stress Level, Sleep Hours, Physical Activity, Diet Score, Smoking, Alcohol, and Family History. The target variable is Medical Condition which categorizes individuals into health states such as Diabetes, Hypertension, Obesity, Asthma, or Healthy.

3.2 Data Preprocessing

The data was pre-processed on a larger scale prior to any model application so that predictions would be trustworthy. First detected and handled the missing/null values using imputation methods like mean, or mode depending on attribute type [7]. Duplicated and irrelevant information were removed for consistency. Outliers were detected using visualizing methods, including box plots and statistical Z scores, then normalization was applied to make the features ranges for glucose, cholesterol and BMI matched [8]. Categorical variables (such as gender, smoking and alcohol) were converted from labels to numeric data in order for them to be read by ML algorithms. The cleaned data set was then divided into training and testing sets in an 80:20 ratio to objectively assess the performance of the model.

3.3 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis was performed for the relationships among various health indicators and also to detect the most influential features of prediction. Fig. 1 shows correlation heatmaps, pair plots and distribution graphs have been built to be able to visualize the relationships among blood pressure, glucose, body mass index (BMI) and HbA1c. From analysis, it was found that high values of glucose and BMI are most predominant to diabetes and stress and cholesterol were most sensitive parameters for hypertension and obesity. This process has the function to improve feature selection by removing redundancy and therefore optimize model efficiency.

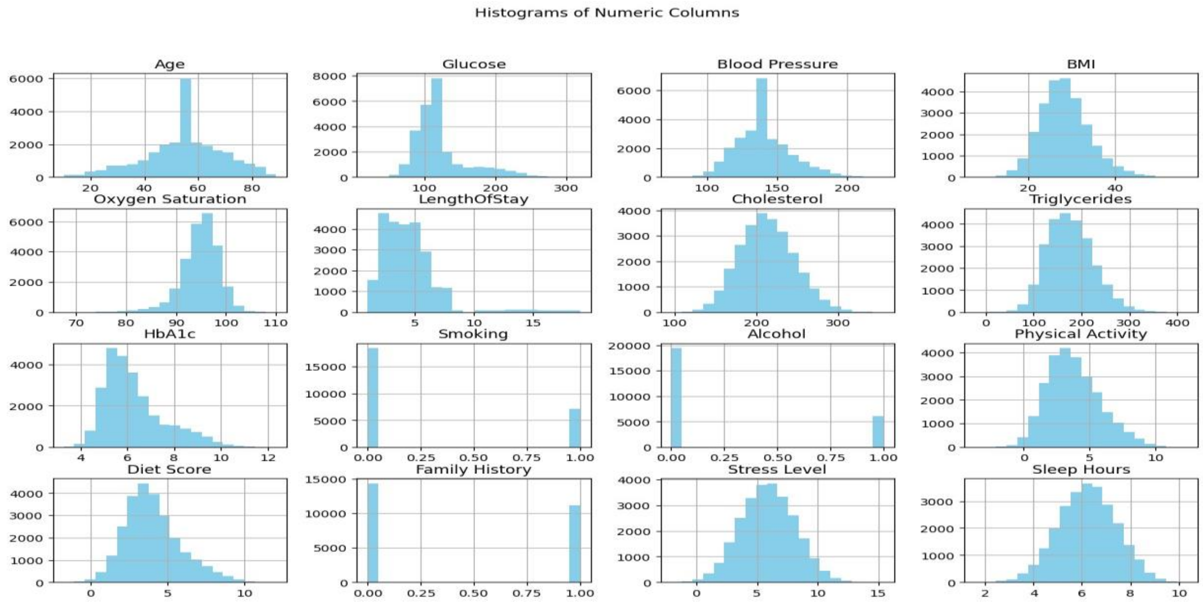

Fig. 1. Histogram representation of data analysis

3.4 Modelling techniques adopted and Evaluation criteria

The pre-processed data set was divided into training and testing sets to objectively verify the performance of models. The relationship between the independent and target output was learnt using machine learning approaches. Performance of the prediction was assessed in terms of standard evaluation measures accuracy, F1-score and ROC curve. Cross-validation principles were applied to guarantee that the trained models are reliable and generalizable [9]. The proposed methodology offers a standardized and repeatable way to build AI predictive models aiming at supporting clinical decisions based on data-derived evidence

Fig. 2 represents the flowchart of the process. The framework used in this work is organized in the way that the dataset goes through preprocessing and feature extraction before feeding Machine Learning. Logistic Regression is used to determine linear relationship between independent health variables and the target variable. Nonlinear dependencies, and feature hierarchies: Decision Tree algorithms such as the Random Forest apply a non-linear approach to both inform the construction of new explanatory indices and to strengthen model robustness. The K-Nearest Neighbors (KNN) algorithm is also a distance-based learner for local similarity of data. All the algorithms add distinctive aspects to enhance prediction performance and further understanding of the medical relevance.

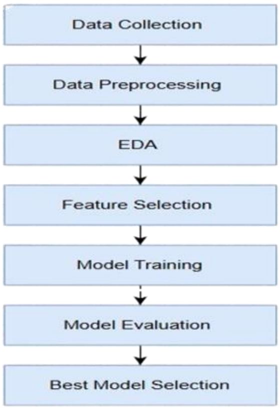

Fig. 2. Flow diagram of the process

4. PROPOSED SYSTEM

Fig. 3 presents the system architecture of the proposed medical condition prediction model including several interconnected modules to address data acquisition, preprocessing, model training and prediction generation, accordingly. The system commences with the acquisition of data that includes the patient's information and healthcare parameters. The preprocessing module includes data cleaning, normalization and encoding required for the dataset analysis [10]. The processed data is then input into a variety of machine learning models including Logistic Regression, Decision Tree, Random Forest, and KNN to train and evaluate predictive accuracy. The predicted disease and model performance is outputted by the output module.

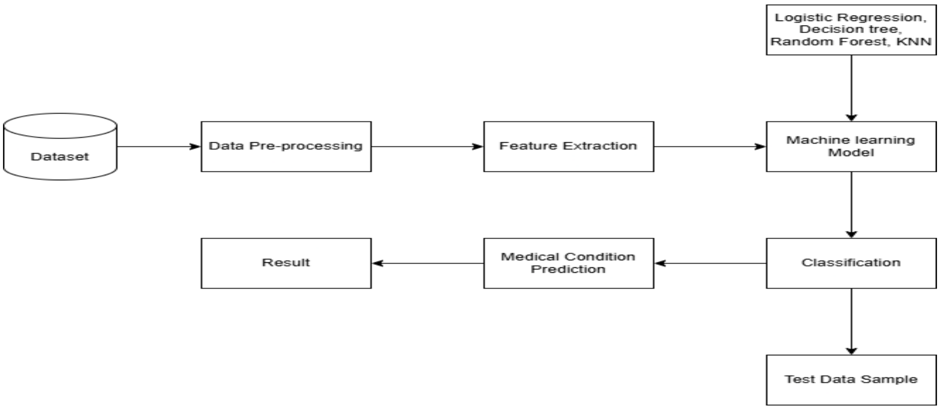

Fig. 3. Workflow for the proposed system

4.1 Modelling techniques used for comparison

1. Logistic Regression was used to capture linear dependencies between health features and achieved solid baseline accuracy.
2. Decision Tree was used because it is easy to model high feature partitions that are easily readable by classification rules.
3. Random Forest is used to improve the generalization and to cope with nonlinear dependencies (a set of decision trees).
4. K-Nearest Neighbors (KNN) predicted the labels of the samples considering their similarity and acted on instance.

4.2 Evaluating the Model Used

The process starts with data input and automatic processing of null or inconsistent features in the patient’s medical records. Feature extraction identifies those health indicators that have a higher impact on a patient's certain diseases. Model training and validation are performed using the four selected algorithms Logistic Regression, Decision Tree, Random

Forest and KNN. The performance of the models is compared after evaluation with major metrics including accuracy, F1-score and ROC-AUC value. The end result of accuracy can be seen in Fig. 4, it is the relative accuracy of the models, and it can be observed that Random Forest and Logistic Regression have an accuracy of 0.91, then Decision Tree with an accurate value of 0.84 and finally KNN with an accuracy of 0.81.

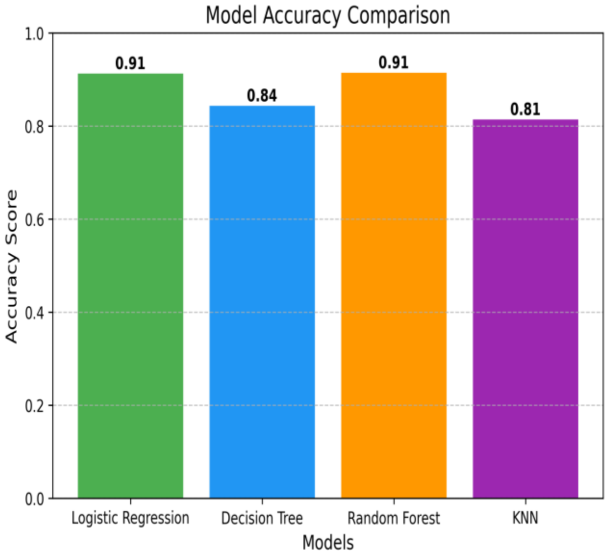

Fig. 4. Model accuracy comparison

This is more of a measure that combines how well precision and recall complements each other. It can take values between 0 & 1 and tells us how well the test precision matches with the recall. It is otherwise called the harmonic mean of precision and recall which encapsulates false positive and false negative errors in one score. In this work, the F1-score is employed to evaluate how well these models can predict different diseases.

Fig. 5 represents bar chart of the F1 score macro average picked from the testing set (Logistic Regression, Decision Tree, Random Forest and K-Nearest Neighbor), Random Forest is the best model among them with an F1 score of approximately 0.90 which demonstrates highest precision and recall trade-off. The next best performing model is Decision Tree (F1 Score of = 0.83) followed by Logistic Regression and KNN (F1 score = 0.78). As such, Random Forest is the best one in this classification task on the basis of macro F1 score (F1 Score =0.90).

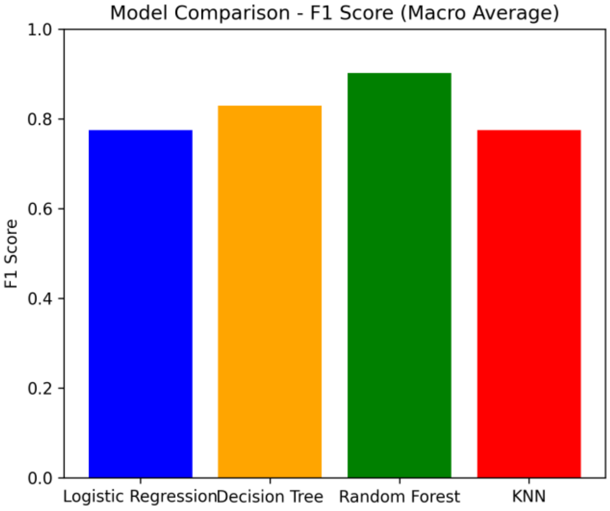

Fig. 5. F1 Score comparison

The ROC curve (Receiver Operating Characteristic), illustrates how well a binary classification test is able to discriminate between two classes. It is the plot of True Positive Rate (Sensitivity) on Y-axis versus False positive rate(1- Specificity)) on X –axis, across all possible thresholds. In this project, we use the ROC curve to evaluate the performance of varying machine learning models based on how well one is able to predict medical diagnosis.

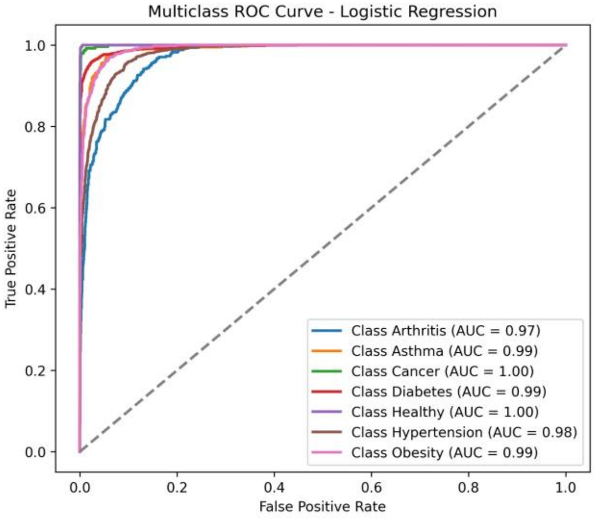

Fig. 6. Multiclass ROC curve of logistic regression model

We can clearly see from Fig. 6 and Fig. 9 that both the ROC curves for Random Forest and Logistic Regression models achieve the highest classification performance across all disease classes with almost perfect Area Under Curve (AUC) close or equal to 1.0 for all the categories. We certainly observe that Random Forest is the most robust, followed precisely by Logistic Regression for each class in both cases we get quite good separation between true and false positives. KNN (shown in Fig. 7) shows a bit lower AUCs, which is the sign of not-so-good performance especially Arthritis and Cancer classes, while Decision Tree also does have in general the lowest AUC values meaning unreliable separation capability for certain classes, especially Arthritis. In conclusion, We were able to find the best suitable and rapid model for our multiclass medical predictions whose results are: The highest one is Random Forest while the lowest one is Decision Tree (shown in Fig. 8) from most accurate to less.

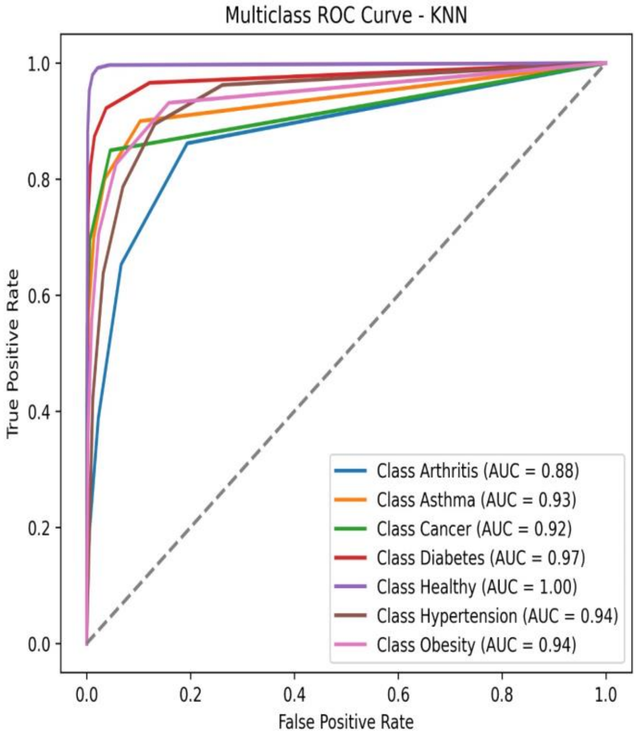

Fig. 7. Multiclass ROC curve of KNN model

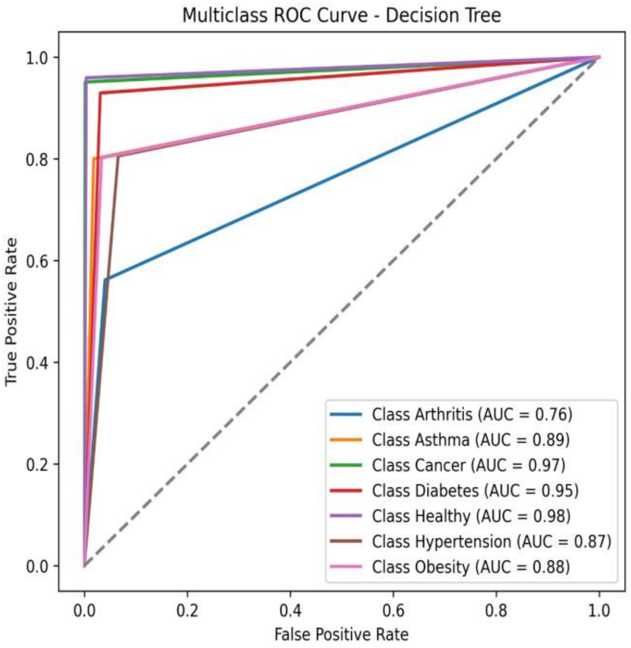

Fig. 8. Multiclass ROC curve of decision tree model

5. Results & Discussions

Table 1 demonstrates the comparative studies of four machine learning models: Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN) demonstrated that ensemble-learning played well on predicting multiple medical condition. Among them, by F1-score, the Random Forest model performed the best (F1 score = 0.90).

The accuracy of models was observed in the test set, it is order of better performer: Random Forest 91.63%, Logistic Regression 91.37%, Decision Tree 84.16% and KNN had presented accuracy of 81.53%. The ROC-AUC analysis additionally indicated the superiority of RF over majority of disease categories, reporting almost perfect AUC for Cancer diagnoses (1.00), Healthy controls (1.00), Asthma cases (0.99), Diabetes patients (0.99) and Obese subjects (0.99). In contrast, Decision Tree showed less AUC value even for Arthritis (0.76) and Hypertension (0.87), indicating less fit with reduced generalization, possibly overfitting.

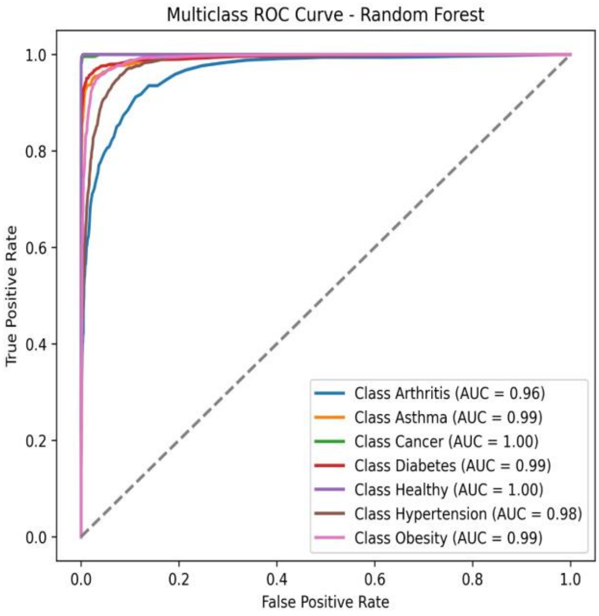

Fig. 9. Multiclass roc curve of random forest model

Parameter	**Logistic Regression**	**Decision Tree**	**KNN**	**Random Forest**
Accuracy	91.37%	84.16%	81.53 %	91.63 %
F1 Score	0.78	0.83	0.78	0.90
Class Arthritis (AUC)	0.97	0.76	0.88	0.96
Class Asthma (AUC)	0.99	0.89	0.93	0.99
Class Cancer (AUC)	1.00	0.97	0.92	1.00
Class Diabetes (AUC)	0.99	0.95	0.97	0.99
Class Healthy (AUC)	1.00	0.98	1.00	1.00
Class Hypertension (AUC)	0.98	0.87	0.94	0.98
Class Obesity (AUC)	0.99	0.88	0.94	0.99

TABLE I. PERFORMANCE COMPARISON OF ALL MACHINE LEARNING MODELS FOR DIFFERENT MEDICAL CONDITION CLASSES

6. CONCLUSION

This paper introduced a medical condition prediction system based on machine learning with the perspective of helping in the early identification and preventative health care. The random forest classifier also had the highest predictive capability and the highest accuracy of 91.63 and the highest F1-score of 0.90 based on comparative analysis. It was also proved to be superior because the values of ROC-AUC were closely similar to 1.00 in most clinical conditions that are significant such as cancer, healthy condition, asthma, diabetes and obesity which implies that it is very robust with noisy clinical data and that its overfitting susceptibility is also low. The average accuracy and ROC-AUC performance of the Logistic Regression were also high and the Decision Tree models were less accurate and had lower F1-scores and greater sensitivity to the class imbalance in the case of such diseases as arthritis and hypertension. The observation that none of the models overfit or underfit the data implies that the models have a high level of generalizability to unseen data. The findings testify to the fact that analytical instruments can be used in the interpretation of the significant clinical and lifestyle factors including glucose level, blood pressure, BMI, cholesterol, and behavioral indicators to be regarded as a valuable addition to a traditional medical assessment. The proposed predictive system will enable clinicians to have a better aptitude to identify the risks as soon as possible, reduce the instances of diagnosis errors, and simplify the process of planning the individual treatment.

REFERENCES

[1] A. H. Abdalla, *Healthcare Risk Factors Dataset*. Kaggle, 2025. [Online]. Available: https://www.kaggle.com/datasets/abdallaahmed77/healthcare-risk-factors-dataset/data

[2] B. Zhang, H. Shi, and H. Wang, "Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach," *J. Multidiscip. Healthcare*, vol. 16, pp. 1779–1791, 2023, doi: 10.2147/JMDH.S410301.

[3] I. H. Sarker, "Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective," *SN Comput. Sci.*, vol. 2, no. 5, p. 377, 2021.

[4] S. R. Sitaraman, "AI-driven healthcare systems enhanced by advanced data analytics and mobile computing," *Int. J. Inf. Technol. Comput. Eng.*, 2025.

[5] H. K. Bharadwaj *et al.*, "A review on the role of machine learning in enabling IoT-based healthcare applications," *IEEE Access*, vol. 9, pp. 38859–38890, 2021.

[6] P. Tanwar, T. Kumar, K. Kalaiselvi, H. Raza, and S. Rawat, *Predictive Data Modelling for Biomedical Data and Imaging*. CRC Press, 2024.

[7] A. Rahman *et al.*, “Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities,” *AIMS Public Health*, vol. 11, no. 1, p. 58, 2024.

[8] E. Kriegova, M. Kudelka, M. Radvansky, and J. Gallo, “A theoretical model of health management using data-driven decision-making: the future of precision medicine and health,” *J. Transl. Med.*, vol. 19, no. 1, p. 68, 2021.

[9] K. N. Qureshi, S. Din, G. Jeon, and F. Piccialli, “An accurate and dynamic predictive model for a smart M-Health system using machine learning,” *Inf. Sci.*, vol. 538, pp. 486–502, 2020.

[10] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, “Applications of machine learning predictive models in chronic disease diagnosis,” *J. Pers. Med.*, vol. 10, no. 2, p. 21, 2020.