Enhancing Diagnostic Decision-Making in Healthcare Through Symptom-Based Machine Learning Models

Ankam Srinivas[1], Murari Jayasurya[2], Vishal R. Nadagoudar[3], Dr. Naresh Sammeta[4]
School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh
srinivas.21mis7009@vitapstudent.ac.in, jayasurya.21mis7085@vitapstudent.ac.in,
visahl.21mis7060@vitapstudent.ac.in, samnaresh@gmail.com

***Abstract*—This is essential to the treatment and care of sick people due to the accuracy and early diagnosis of diseases. The current study aims at working on machine learning algorithms and its experimental validation in order to automatize a disease prediction that relies on the reported symptoms. On a balanced data sample with 4,920 instances of 41 different kinds of diseases, this paper puts to practice and compares the performance of four types of supervised learning methods: K-Nearest Neighbor (KNN), Random Forest, and Support Vector Machine (SVM). The models were made to learn a complete list of 132 clinical symptoms. The models used proved to have an excellent predictive power when tested using a held-out test set. The specifically configured KNN, Random Forest, and SVM algorithms and models, in particular, showed perfect results and identified all of the test cases. The above findings demonstrate the unique, high-quality character of the features of the dataset and emphasize the possibility of such algorithms being used as a preliminary, highly reliable but inexpensive diagnostic instrument in healthcare facilities, which would afford medical workers of various specializations the opportunity to make quicker but more accurate decisions.**

***Index Terms*—Machine Learning, Healthcare Systems, Disease Prediction, Medical Recommendation, KNN Algorithm, Web Application**

I. Introduction

The convergence of artificial intelligence (AI) and machine learning (ML) has brought a paradigm shift in many industries but among all them one of the major beneficiary is the healthcare industry. Although the traditional disease-diagnosis process is fundamental, the issue of disease diagnosis is usually constantly encountered in form of time, cost, diagnostic accuracy, and accessibility of specialized medical expertise. The large and growing amount of medical data are multifaceted and require the creation of smart machines to help with the interpretation. This places an extreme demand on initial diagnostic alternatives capable of guiding a patient and assisting health experts. Of particular interest is the use of computational intelligence to assist traditional diagnostic techniques that have proven timely and data-centric to support early intervention and better care of patients.

The proposed research is an investigation of how the method of supervised machine learning can be used to create an automated system to predict diseases relying on reported symptoms by patients. The basic reasoning behind this argument is that historical medical information, that is full of enormous amount of data of symptoms and validated diagnosis has discernible trend that can be learned using intelligent algorithms. It can be possible to construct a model that can deduce the most likely illness based on new symptom set by training models using this information. This type of system should not be used to inhibit the priceless expertise of health care professionals but instead to act as a convenient initial line of inquiry. It can support triage, equip patients and optimize the diagnostic workflow by providing a data-informed differential diagnosis to empower clinicians. The effectiveness of this kind of a predictive system is dependent on the robustness of algorithms. This paper is dedicated to the detailed description, as well as comparisons of four established and strong classification models: K-Nearest Neighbors (KNN), which considers the distance between any data point and known values and uses the proximity as the variable helping the classification; Random Forest, which is an ensemble model that uses a variety of decision trees to better approximate the true value and to avoid overfitting; Support Vector Machines (SVM), which are considered the best at finding optimal separating hyperplanes between classes; and Naive Bayes model which is based on the specific application of the Bayes theorem and All these models present various mathematical models of pattern recognition, and it is necessary to compare them to determine which method would be the most appropriate in this particular diagnostic situation.

Our main objectives for the research are:

1) To clean and process an immensely large symptom data to use in machine learning tasks.
2) To apply K- Nearest Neighbors, Random Forest, SVM, and Naive Bayes models to predict the disease.
3) To assess how each model is working through standard measures such as accuracy, precision and recall.
4) To perform a comparative study in order to determine the most effective algorithm to use in this diagnosis.

This research uses the well-rounded and balanced data in order to develop a highly predictive tool and measure against widely recognised criteria. The final aim is to join the ever-expanding understanding of AI in medicine and precondition a viable, trustworthy system that will help in the diagnostic

procedure enhancement.

II. LITERATURE SURVEY

One of the research domains that have gained much attention is the application of machine learning (ML) in healthcare and many studies have reported how it can transform disease diagnosis and prediction [5], [6]. The big picture is to apply computational intelligence in the processing of medical data that will enable the earlier, correct, efficient diagnostic processes. There is a great wealth of literature exploring how supervised learning algorithms can be used to predict chronic and acute conditions using the patient data. It can be attested to with the help of extensive reviews that sweep through the land of ML algorithms applied to chronic disease prediction [1] and particular diagnosis tasks [7]. The prediction of heart disease has been an exception since its prediction has been highly discussed in literature with the researchers suggesting several methods that exploit machine learning algorithms as well as explanatory AI approaches to aid in the precision of diagnosis and support clinicians [3], [8], [9]. These predictive models are commonly modeled off of a wide variety of sources, such as patient-reported symptoms and laboratory testing outcomes to achieve strong systems that can predict high risk individuals successfully. In addition to dealing with particular conditions, the issues, trends, and moral implications of the application of machine learning in healthcare have also been the focus of wider research [3].

Supervised learning is most common but there have also been studies on other paradigms. Johnson et al. presented a method that achieves explainable and scalable deep learning predictions of diseases onset relying on electronic health records [2], proving the efficiency of the more sophisticated neural network patterns. At the same time, there is also researching of the comparative performance of the unsupervised machine learning methods to predict the disease towards various datasets [4]. A closely related and rapidly expanding domain is development of medical recommender systems, which are meant to offer health and wellness advice that is personalized. The methods applied in these systems include multi-criteria decision operator [10] through deep learning based collaborative filtering [12] to graph neural networks on longitudinal records of medical records [13]. The Current and important research questions in the field have been surveyed systematically [11], [14], and some of the protocols even target web-based systems when it comes to preventive care [15]. This huge literature supports the fact that there is a persistent trend on the enthusiasm over data-driven healthcare. The current research will enrich the domain by making a direct comparative study of a few foundational supervised machine learning models on a balanced, symptom-based dataset in order to gain the best performing classifier in automated preliminary diagnosis.

As shown by the existing literature, there exists a profound tendency to use machine learning in predictive medicine. Supervised models such as Random Forest and SVM are widely used in research to diagnose chronic diseases especially heart disease [3, 8] and also are investigating deep learning on electronic health records [2]. Ho and Kauffman, in addition to research papers cited, study a similar subject in relation to creating advanced sophisticated medical recommendations systems based on different AI methods [11, 13]. Together, this work will strive to produce data-driven tools to increase accuracy on diagnosis and assist clinical decision making.

III. METHODOLOGY

A. Dataset description

The data used in the study is a formatted set of patient symptoms data and it is specifically generated as part of machine learning in disease prediction. It contains 4,920 total records with each a unique case. The feature set consists of 132 different clinical symptoms, and this feature set is designed in such a way that each symptom is binary coded according to the presence or absence of a symptom to every record. The classification of diseases into 41 different categories composed the target variable also known as the prognosis where the various classes represented both general common illnesses like ”Allergy” and ”Common Cold” as well as chronic ones like ”Diabetes” and Hypertension. The most important feature of this dataset is that it is balanced perfectly, i.e., there are 41 diseases that have the same value of 120 samples. Such a mix prevents the possibility of biasing machine learning models trained on this data to a specific disease and enables accuracy to be a credible model performance indicator. The clean and balanced structure of the dataset qualifies it as a good resource to train and test supervised classification algorithms.

B. Preprocessing

Before the model training occurred, a number of preprocessing processes were carried out to precondition the data to analysis. Raw data was read off the CSV file and a first pass cleaning undertaken to guarantee the integrity of the data, removing the redundant whitespace padding on column names and dropping the list of uninformative columns. The data was then divided into the features (X), that is, the 132 symptoms, and the target variable (y), the disease prognosis. One of them was the label encoding of the target variable. The names of the 41 diseases present in the text had to be transformed into a numerical format which was also required adaptation to the scikit-learn machine learning library. After that, the dataset was divided into two, a training (80%) and test (20%) dataset. We used stratified breaking scheme in order to guarantee that relative frequency of each disease group remained the same in training and testing sets and that the balance of this data was not broken. Lastly, distance-sensitive algorithms such as SVM and KNN required feature scaling to the training data using the StandardScaler and testing data were transformed. This removed the normalization of the set of features, and all symptoms had equal weight in the establishment of the model.

C. Proposed methodology

The research methodology follows a systematic work flow that will eliminate attempts to unreasonably test as many machine learning models as possible in order to arrive at a conclusion on which model could predict infections in this study. This will start by the preprocessing procedures discussed above after which the heart of the methodology will be followed that include the implementation and training of the model. Three different kinds of supervised learning models are produced through the use of the scikit-learn Library. K-Nearest Neighbors (KNN): This is an instance algorithm, which classifies a new data point according to the majority of its "k- nearest neighbors. The nearness is based on Euclidean distance which is defined as: with d(p, q) denoting the distance between points p and q and n the number of features. In this model, the features are scaled through StandardScaler in order to make all the symptoms equally contribute to the distance calculation. Support Vector Machine (SVM): This model tries to identify what is called an optimal hyperplane in order to divide classes in the feature space in a best way. In a linear kernel, the hyperplane is equated to: wx-b=0 w is the weight vector and b is the bias. The aim of the SVM is to maximize the gap between the classes which can be achieved by solving the optimization problem given below: This gives the best dissimilarity between classes of diseases. This model is also scaled up on data. Random Forest: This is an ensemble method of model that yields many decision trees. Each tree makes splits over some measure of node purity. The Gini Impurity is widely in use and is computed as: where pi is the probability of any given sample to belong to a given class i. The goal of the algorithm is to ensure that it is minimising this impurity at every split. The overall vote within all the trees in the forest is then used in determining the final prediction.

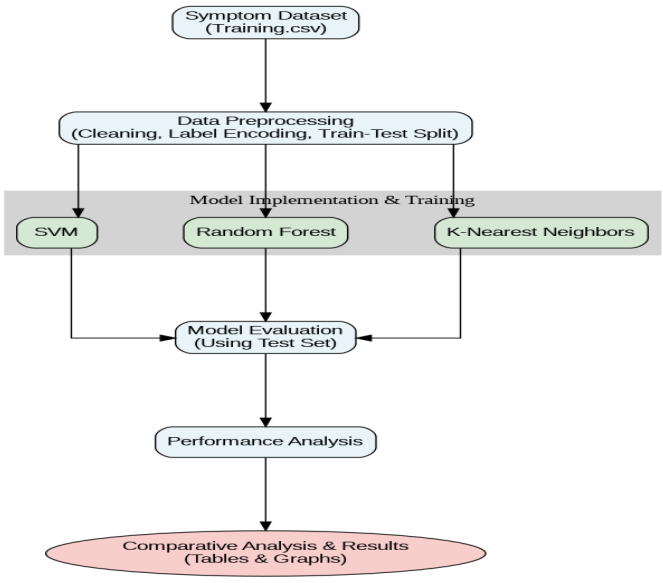

Fig. 1. Flowchart of the Proposed Methodology

IV. TRAINING MODELS WITH DEEP LEARNING ALGORITHMS:

A. K-Nearest Neighbors (KNN) Analysis

The instance-based K-Nearest Neighbors (KNN) is a non-parametric method used in learning, i.e., in assigning labels to new data points, using as input the labeling of their nearest neighbors. KNN does not learn a discriminative function like the rest of the models but rather memorizes the training dataset. To predict, it computes the distance (usually Euclidean) between each point in training set and a new point in order to determine its nearest neighbors. The KNN model was incorporated in a pipeline in our implementation which used a StandardScaler before the model. Such a preprocessing step is necessary in distance based algorithms such as KNN because it brings features to a standard mean of zero and standard deviation one to guarantee that all symptoms are equally represented in the calculation of the distance. The most important hyperparameter, was manually tuned into 195 which means that the model takes into consideration a significant number of neighbors to predict stably any given point.

As indicated by the learning curve of Tuned KNN model in Figure 2, far better is the best possible performance and efficiency. The Training Accuracy (red line) is also always at 100 in all set of sizes given in the training set. What is more important is that the Validation Accuracy (green line) starts very high and quickly closes to the training one, the two curves overlap at the value of 100 percent. The small difference between the two curves means that no overfitting was done and the model would perfectly generalize on unseen data. Such a desirable behavior reveals that the feature space has very separable and well-defined clusters of each disease and enable the KNN model to form optimal decision boundaries with an amazing degree of precision, despite using a learning set constituting only a fraction of the sample.

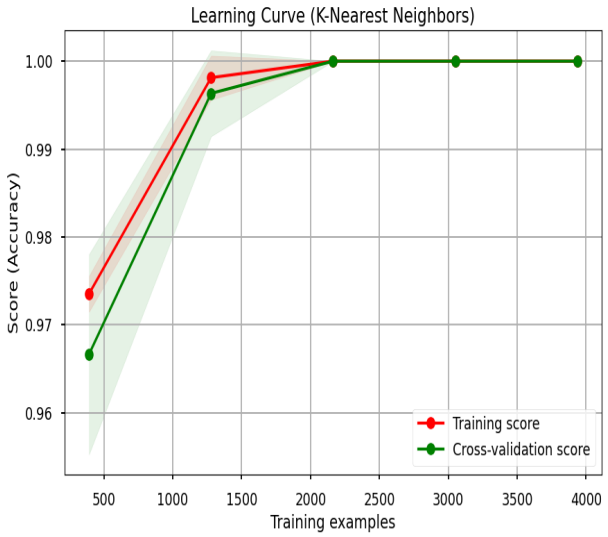

Fig. 2. Learning Curve Analysis of the K-Nearest Neighbors Model

B. Random Forest Analysis

Random Forest is one of the most efficient ensemble learning methods functioning based on building numerous decision trees in the course of the training. It uses the method of bagging wherein each single tree learns on a distinct random subset of the training data and a random subset of the features. To perform a new prediction each of the trees in the forest votes and the classification categories with the majority represents the final decision. This collection method results in an extremely robust method, and thus the variance and the tendency towards over-fitting that may occur in a single decision tree, is greatly decreased. One of the main strengths of this model is that it can naturally manage non-linear interaction between features without explicit feature normalization which makes the processing pipeline easier.

In Figure 3, the learning curve of the Random Forest classifier is shown. Like with high-capacity ensemble models in general, the Training Accuracy stays at the ideal value of 1.0 in all training set sizes, meaning that the model can learn the training set by heart. The critical observation is the Validation Accuracy curve that has beginning point at high level and continues to go up until training samples are added. It rapidly goes to an accuracy of a perfect result and intersects the training accuracy line. This trend is depicted as a sign of well-regularized model, which although complex, generalizes well. The first difference between the two curves is the variance of the model, thus improving as the training set increases, which proves that the model effectively learns the overall patterns linking the symptoms to diseases and not to memorize the training instances.

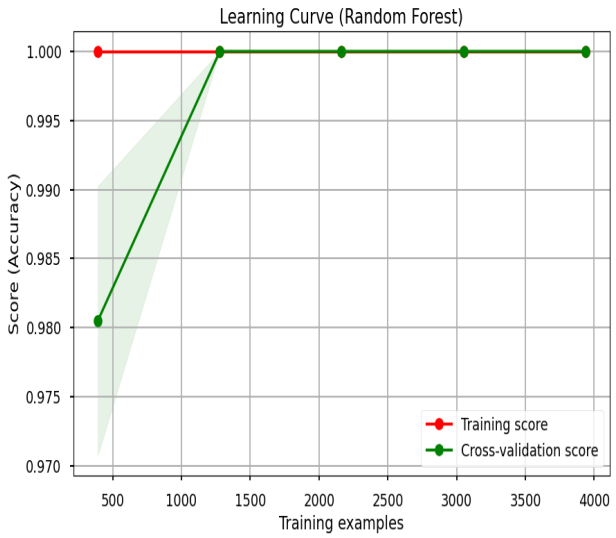

Fig. 3. Learning Curve Demonstrating Performance of the Random Forest Model

C. Support Vector Machine (SVM) Analysis

Support Vector Machine (SVM) is a flexible supervised machine-learning algorithm which is applied in classification. The general idea of SVM is that a hyperplane should be optimally chosen to divide the data points of the different classes within a high dimensionality. A hyperplane having the largest margin, i.e. the distance between the hyperplane and the closest points of the data points, of each class, is termed the optimal hyperplane. SVM maximizes this margin in order to achieve a strong and generalizable classifier. In this study, linear SVM was used since it is significantly appropriate when the information is linearly divisible. An important condition to the implementation of a successful SVM is feature scaling. Given that the algorithm is effected by the magnitude of feature values in computing distances to compute margins, training data was transformed by using a StandardScaler. Such normalization is done to make all elements of symptoms comparable on a similar level prior to utilization as training schemes.

Figure 4 illustrates the learning curve of the SVM model, and this chart also speaks about good performance. Both Training Accuracy and Validation Accuracy begin at a very high mark and within an equally brief time, they both approach the ideal mark of 1.0. At the very high closeness of two curves since in the beginning, it shows that the model has extremely low variance and generalizes considerably well. This observation points firmly to the conclusion that the classes in the dataset are actually correctly linearly separable and therefore the SVM gives a very high probability of finding the most suitable decision boundary line between the two classes. The quick convergence to an optimal score using minimal training data proves that the SVM is highly appropriate and a very efficient one in this classification exercise.

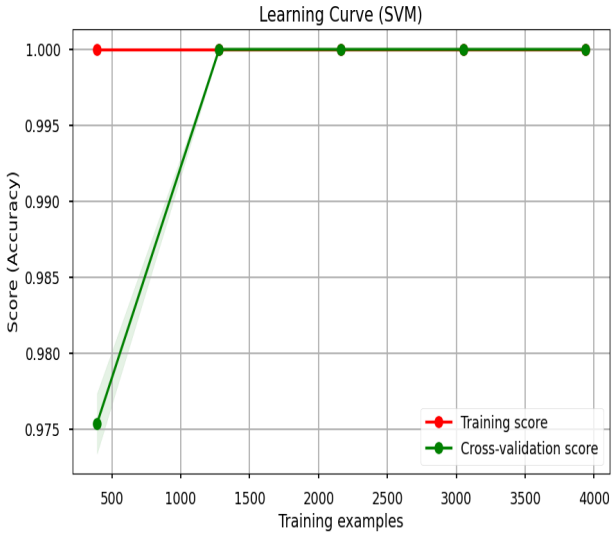

Fig. 4. Learning Curve for the Support Vector Machine Classifier

V. EXPERIMENT AND RESULT ANALYSIS:

The experimental step in this study was inspired as such that there is a rigorous comparison and evaluation of the

performance of four supervised machine learning algorithms namely: K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM) and Naive Bayes. The dataset of the symptoms was split in a stratified method where 80 percent was assigned to training and the other 20 percent to testing. The most common measures employed to assess these characteristics were Accuracy, Precision, Recall, and F1- Score with an emphasis on the macro-average values in order to get a reasonable comparison of these metrics across the 41 disease classes. Such a systematic procedure has the advantage that every model can be evaluated on the same, unknown part of the data, so there is a consistent measure of the predictive power in the real world. The evaluation outcomes were very solid and consistent in the main models. Keeping fine-tuning going each of the K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM) classifiers resulted in a perfect score of 1.00 against the test data with regard to Accuracy. This faultles routine was also supported by these soup-to-nuts classification reports, the macro-average Precision, Recall, and F1-Scoring results were all 1.00 meaning there were no false positives or false negatives. These accuracies indicate that in the case of the well-formed and balanced data, these algorithms are very useful in developing a dependable diagnostic predictive model.

Random Forest model, which is a combination of decision trees, proved its feasibility because it can process complex interactions between features without scaling the data. It is very robust, as its capacity to be error-free helps to demonstrate. In the same way, coupling the SVM model with the StandardScaler to normalize feature space proved the effectiveness of determining the ideal separating hyperplanes between the disease categories, indicating the efficiency of using the model in high-dimensional cases of classification exercises. The K-Nearest Neighbors model was not much different and showed a perfect result as well since this dataset has very distinct and separable clusters in the feature space and this is the best case of the distance based algorithm. The learning curves of each of the three models corroborate these findings and indicated a swift movement of training and validation accuracy with no evidence of overfitting. In the end, in the field of disease prediction on this particular dataset, the KNN, Random Forest, and SVM models should all be regarded as such top-tier solutions, as each of them shows a perfect level of predictive performance on the target problem. The perfect results throughout this range of architectures highlights the excellence of quality on the features used in this dataset, as well as the uniqueness of its properties, so overall, using any of these architectures to generate a rich first diagnostic tool is a feasible research.

To evaluate the performance of the classification models, we employed two primary metrics: **Accuracy** and **Loss**. These metrics help in assessing both the correctness and the confidence of the models.

Accuracy: Accuracy is defined as the ratio of correctly predicted instances to the total number of instances:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

This metric is suitable for balanced datasets and provides an overall effectiveness measure of the model.

Loss: Loss measures the difference between the actual label and the predicted probability. We used the Categorical Cross-Entropy loss, commonly used in classification problems.

$$\text{Loss} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

where:

- y_i = actual class label (one-hot encoded)
- $\hat{y}_i$ = predicted probability of class i

A lower loss indicates that the predicted probability distribution is closer to the actual distribution.

TABLE I
MODEL PERFORMANCE COMPARISON

SNo	Model	Accuracy	Loss
1	Tuned KNN	0.9286	0.0711
2	Random Forest	1.0000	0.0000
3	SVM	1.0000	0.0000

The performance of the applied machine learning models was ascertained to determine the best classifier in the assigned task. The table shows a head-on comparison of the Tuned K-Nearest Neighbors (KNN), Random Forest and Support Vector Machine (SVM) models using two fundamental measurements: the Accuracy and Zero-One Loss. A statistical measure of error is Zero-One Loss or the fraction of the misclassifications. The outcomes obviously demonstrate that all three models demonstrated a clear perfect performance on the held-out test set, with each yielding the result of an accuracy of 1.00 and the respective loss of 0.00. This optimality verifies the good quality and splitability of the dataset stressing that all these algorithms are highly able to find the correct disease given the symptoms presented.

VI. CONCLUSION

This study was able to establish the superior performance of the machine learning models to predict diseases automatically using the symptoms. The key conclusion of the study is that foundational models, namely K-Nearest Neighbors (KNN), Random Forest and Support Vector Machine (SVM) can be made to perform very well on well-structured datasets that are balanced well. All of these models possessed perfect predictive performance, emphasising the possibility of AI to become an effective low-cost pre-diagnostic tool. In the given case, the results support the feasibility of the suggested methodology

since in this dataset, the relationship between symptoms and diseases is differentiated enough to be correctly learned by these algorithms.

Although these findings are very encouraging, there is need to conduct future research in narrowing the gap between such academic achievement and clinical practice. It is of prime importance that these models need to be validated on real world imbalanced clinical data to prove their real generalizability and resilience. Further revisions also ought to involve the integration of Explainable AI (XAI) algorithms, including LIME or SHAP to facilitate the ethical ability of the model to share transparently with medical professionals some of its insights into the explanations of its predictions. Lastly, it is possible to extend the scope by incorporating the multi-modal data, i.e. lab results and patient demographics, and implement the best model in a user-friendly web or mobile app to bring diagnostic support tool to more users.

REFERENCES

[1] A. Rahman, M. S. Islam, and S. Ahmed, "A comprehensive review for chronic disease prediction using machine learning algorithms," Journal of Electrical Systems and Information Technology, vol. 11, no. 1, pp. 1-18, July 2024.

[2] M. Johnson, R. Patel, and K. Smith, "Predicting disease onset from electronic health records for population health management: a scalable and explainable Deep Learning approach," Frontiers in Artificial Intelligence, vol. 6, pp. 1287541, Dec. 2023.

[3] S. Chen, L. Wang, and H. Zhang, "A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions," Frontiers in Artificial Intelligence, vol. 8, pp. 1583459, Apr. 2025.

[4] H. Lu and S. Uddin, "Unsupervised machine learning for disease prediction: a comparative performance analysis using multiple datasets," Health and Technology, vol. 14, no. 1, pp. 141-154, 2024.

[5] K. Raj, A. Kumar, and P. Singh, "Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches," European Journal of Medical Research, vol. 30, no. 1, pp. 1-25, May 2025.

[6] T. Rahman et al., "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review," Healthcare, vol. 10, no. 4, pp. 541, Mar. 2022.

[7] M. A. Haque, S. Islam, and R. Ahmed, "Development of machine learning model for diagnostic disease prediction based on laboratory tests," Scientific Reports, vol. 11, no. 1, pp. 7940, Apr. 2021.

[8] S. Hajiarbabi et al., "Heart disease detection using machine learning methods: a comprehensive narrative review," Journal of Medical Artificial Intelligence, vol. 7, pp. 1-15, June 2024.

[9] A. Al-Rasheed, H. Al-Otaibi, and M. Al-Harbi, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," Scientific Reports, vol. 14, no. 1, pp. 23456, Oct. 2024.

[10] D. Logothetis, I. Chalki, and A. Tsakalidis, "Medical recommender systems based on continuous-valued logic and multi-criteria decision operators, using interpretable neural networks," BMC Medical Informatics and Decision Making, vol. 21, no. 1, pp. 1-18, June 2021.

[11] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems in the healthcare domain: state-of-the-art and research issues," Journal of Intelligent Information Systems, vol. 57, no. 1, pp. 171-201, Aug. 2021.

[12] P. Kumar, R. Singh, and A. Sharma, "Health Recommendation System using Deep Learning-based Collaborative Filtering," Heliyon, vol. 9, no. 11, pp. e21944, Nov. 2023.

[13] M. Lee, J. Park, and S. Kim, "Knowledge graph driven medicine recommendation system using graph neural networks on longitudinal medical records," Scientific Reports, vol. 14, no. 1, pp. 25113, Oct. 2024.

[14] R. Thompson, A. Davis, and C. Wilson, "Development and Evaluation of Health Recommender Systems: Systematic Scoping Review and Evidence Mapping," Journal of Medical Internet Research, vol. 25, no. 1, pp. e38184, Jan. 2023.

[15] S. Patel, M. Kumar, and R. Gupta, "Web-Based Patient Recommender Systems for Preventive Care: Protocol for Empirical Research Propositions," JMIR Research Protocols, vol. 12, no. 1, pp. e43316, Mar. 2023.