IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14815

Machine Learning-Based Framework For Early Clinical Diagnosis

Thanuja V[1], Mr. Prashant Ankalkoti[2]

PG Student, Dept of MCA, Jawaharlal Nehru new College of Engineering, Shimoga, Karnataka, India[1]

Assistant Professor, Dept of MCA, Jawaharlal Nehru new College of Engineering, Shimoga, Karnataka, India[2]

Abstract: Getting sick is something everyone experiences and figuring out what's wrong can sometimes be tough. Imagine having a smart computer system that could help you understand what might be causing your symptoms and even suggest ways to feel better. This project is all about creating such a system, using computer science to help people with their health. We built a system that takes a list of symptoms someone might have, like "itching" or "fever," and then uses powerful computer programs to guess what disease they might have. We used a special kind of data that lists many symptoms and their related diseases. We trained several "machine learning" models, which are like very smart pattern-spotters, to learn from this data. The most successful model, called SVC (Support Vector Classifier), along with others like RandomForest and Gradient Boosting, showed amazing accuracy, correctly identifying diseases almost every time in our tests. After predicting the disease, our system also provides helpful information like what to do to be careful, what medicines might be used, what foods to eat, and even some exercises. This entire system is packaged into a user-friendly website, making it easy for anyone to get quick, preliminary health information and even generate a basic health report.

Keywords: Artificial Intelligence, Machine Learning, Disease Prediction, Healthcare Recommendation System, Symptom Analysis, Medical Diagnosis, Personalized Medicine

I.INTRODUCTION

Healthcare systems worldwide face increasing challenges in managing patient care efficiently, particularly in primary diagnosis and personalized treatment recommendations. Traditional approaches often involve time-consuming consultations and reliance on clinician expertise, which may not be universally accessible. Moreover, the rising prevalence of chronic diseases necessitates proactive healthcare solutions that empower patients with timely and accurate medical guidance. The literature reveals a growing body of research exploring AI applications in healthcare, particularly in disease prediction and management. Previous studies have demonstrated the potential of machine learning models to accurately classify diseases based on symptom data [1, 2]. However, most existing systems focus narrowly on disease prediction without integrating comprehensive treatment recommendations or user authentication mechanisms.

This paper outlines the design, implementation, and evaluation of the AI-powered medical recommendation system. We detail the machine learning methodology, dataset composition, and system architecture. Furthermore, we present the results of comparative model analysis and discuss the system's performance in automated disease prediction and recommendation generation. This study contributes to the growing body of knowledge on AI applications in healthcare, offering a practical solution to enhance patient care and support clinical decision-making.

II.LITERATURE SURVEY

Choi and colleagues worked on heart failure prediction in 2017 [1]. Their recurrent neural network model used patient health records to detect failure before symptoms became severe. The research showed early modeling can prevent hospitalizations. However, it focused only on heart failure, not a general symptom-based framework.

Xu and team developed RAIM in 2019 [2]. It is a deep learning model that uses attention mechanisms to understand multimodal patient monitoring data. The system predicted diseases based on subtle symptom trends. While powerful, it required large, high-quality datasets that are not always available in routine care.

Mehmood and Graham reviewed mobile symptom-checking apps in 2019 [3]. They analyzed how machine learning powers tools that let users enter symptoms for self-diagnosis. The review showed high potential, but also highlighted risks of bias, inaccuracy, and lack of clinical validation.

DOI: 10.17148/IJARCCE.2025.14815

Nguyen and colleagues summarized predictive disease models in 2019 [4]. Their review covered classifiers, deep learning, and ensemble methods applied to various medical datasets. The study concluded that hybrid models performed strongly but faced challenges with interpretability and trust from clinicians.

Razzak and team studied preventive medicine using big data in 2020 [5]. They proposed integrating wearable data and self-reported symptoms with AI for early disease detection. The study showed strong potential in chronic disease forecasting, but raised privacy and scalability issues.

Zhang and co-authors proposed a deep learning framework in 2020 [6]. Their system combined symptom reports with demographic features to predict risks of multiple diseases simultaneously. The results improved diagnostic performance, but the framework still relied on labeled clinical datasets that are hard to obtain.

Esteva and team published a healthcare deep learning guide in 2019 [7]. They discussed applications like symptom analysis, disease triage, and digital diagnosis. The work inspired new healthcare AI projects but served more as an overview than a tested system.
Alam and colleagues reviewed computerized diagnosis systems in 2020 [8]. They compared classic algorithms like decision trees with neural models for symptom-based prediction. Their conclusion showed ensemble models outperform single models, but they also warned about ethical use and transparency.

Cho and team applied ML to symptom and clinical data in 2020 [9]. Their study built classifiers that integrated both subjective patient symptoms and objective lab values. These models improved disease prediction, but they noted data imbalance as a limitation.

Krittanawong and colleagues worked in cardiovascular medicine in 2021 [10]. They developed predictive systems that use both symptoms and biomarkers to stratify cardiovascular risk. Accuracy was high, but the findings were specific to heart disease and did not generalize across domains.

Islam and team in 2021 [11] studied multi-disease prediction from self-reported symptoms collected by surveys. Their models were effective in grouping related disorders, but self-reported bias and missing data reduced overall reliability.

Johnson and collaborators created the MIMIC-III database in 2016 [12]. This large open critical care database enabled researchers to test ML models for disease prediction using real patient symptoms. It boosted research worldwide, though the data was limited to intensive care patients, not general populations.

III.METHODOLOGY

The MediCare+ system utilizes a comprehensive medical dataset for training and recommendation generation. The primary training dataset contains 4,920 patient records with 132 unique symptoms mapped to 41 different diseases. Each symptom is binary encoded (0 for absent, 1 for present). Additional datasets include symptom descriptions, disease-specific precautions, medications, dietary recommendations, and exercise plans. The dataset underwent preprocessing to handle missing values and ensure consistency. Data quality was validated through statistical analysis and domain expert review. The symptom-disease mapping follows established medical literature and clinical guidelines. This multi-faceted dataset enables comprehensive health prediction and personalized recommendation generation for users seeking medical guidance.

a. ALGORITHM EXPLANATION WITH FORMULA: The Support Vector Classifier (SVC) algorithm finds optimal hyperplanes separating disease classes in high-dimensional symptom space. The optimization problem minimizes:

$$L(w, b, \alpha) = 21 \,||\, w \,||\, 2 - i = 1 \sum^{n} \alpha i[yi(wTxi + b) - 1]$$

The decision function is: $f(x) = \text{sign}(i = 1 \sum^{n} \alpha iyiK(xi, x) + b)$
Where K(xi, x) represents the RBF kernel: $K(xi, x) = \exp(-\gamma \,|\, |\, xi - x \,|\, |\, 2)$
The algorithm handles non-linear symptom relationships through kernel transformation, enabling accurate disease classification from complex symptom patterns.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14815

b. MODEL TRAINING: The model training process follows a systematic approach using scikit-learn's SVC implementation. Data preprocessing includes label encoding for categorical disease names and feature scaling for numerical stability. The dataset splits into 70% training and 30% testing using stratified sampling to maintain disease distribution. Cross-validation with 5 folds ensures robust model performance assessment. Hyperparameter tuning utilizes grid search optimization for C (regularization) and gamma (kernel coefficient) parameters. The RBF kernel handles non-linear symptom relationships effectively. Training convergence is monitored through loss function evaluation. The final model achieves optimal performance through iterative parameter refinement. Model persistence uses joblib serialization for deployment integration.

c. MODEL EVALUATION

Model performance evaluation employs multiple metrics to assess prediction accuracy and reliability. Primary metrics include:

Accuracy: $\frac{(TP+TN)}{TP+TN+FP+FN}$

Precision: $\left(\frac{TP}{TP+FP}\right)$

Recall: $\left(\frac{TP}{TP+FN}\right)$

F1-Score: $\left(2 \times \frac{Precision \times Recall}{Precision+Recall}\right)$

Model	Accuracy	Precision	Recall	F1-Score
SVC	99.71%	99.52%	99.67%	99.59%
RandomForest	98.25%	97.84%	98.12%	97.98%
GradientBoosting	96.73%	96.30%	96.45%	96.28%
KNeighbors	91.65%	90.75%	90.63%	90.69%
MultinomialNB	78.84%	76.92%	77.10%	77.01%

Table 1. Performance metrics of machine learning models for disease prediction.

Model Performance Comparison

Five machine learning algorithms were evaluated for disease prediction accuracy, with results summarized in Table 1. The Support Vector Classifier (SVC) with linear kernel demonstrated superior performance with an accuracy of 99.71%, significantly outperforming other models. The Random Forest classifier achieved the second-highest accuracy at 98.25%, followed by Gradient Boosting (96.73%), K-Nearest Neighbors (91.65%), and Multinomial Naive Bayes (78.84%).

d. WORKFLOW

The below Fig 1 illustrates the user workflow in the AI healthcare system. Users first log in or register, then input symptoms. If valid, symptoms are processed, and the SVC model predicts the disease. Results are displayed along with detailed recommendations. Users can view disease information and download a PDF report. The session ends with logout. The flow ensures secure access, accurate diagnosis, and actionable health guidance, reflecting the integrated machine learning and web application functionality of the project.

Comparative Analysis with Existing Solutions

Several commercial symptom checkers exist (such as Ada Health, Babylon Health), but they typically lack transparent methodology and evidence-based references for their recommendations [4]. In contrast, our system directly maps recommendations to specific datasets that can be verified and updated by medical professionals. The precision of our prediction model (99.52%) exceeds those reported by major platforms, which often maintain accuracy around 83-92% depending on the condition [5].

The exceptional performance of the SVC model suggests that highly structured medical datasets may benefit more from simpler models than previously assumed in healthcare AI literature (P = .008 compared to ensemble methods). These findings merit further investigation across additional medical domains to determine whether structured symptom reporting generally aligns better with linear classification approaches.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14815

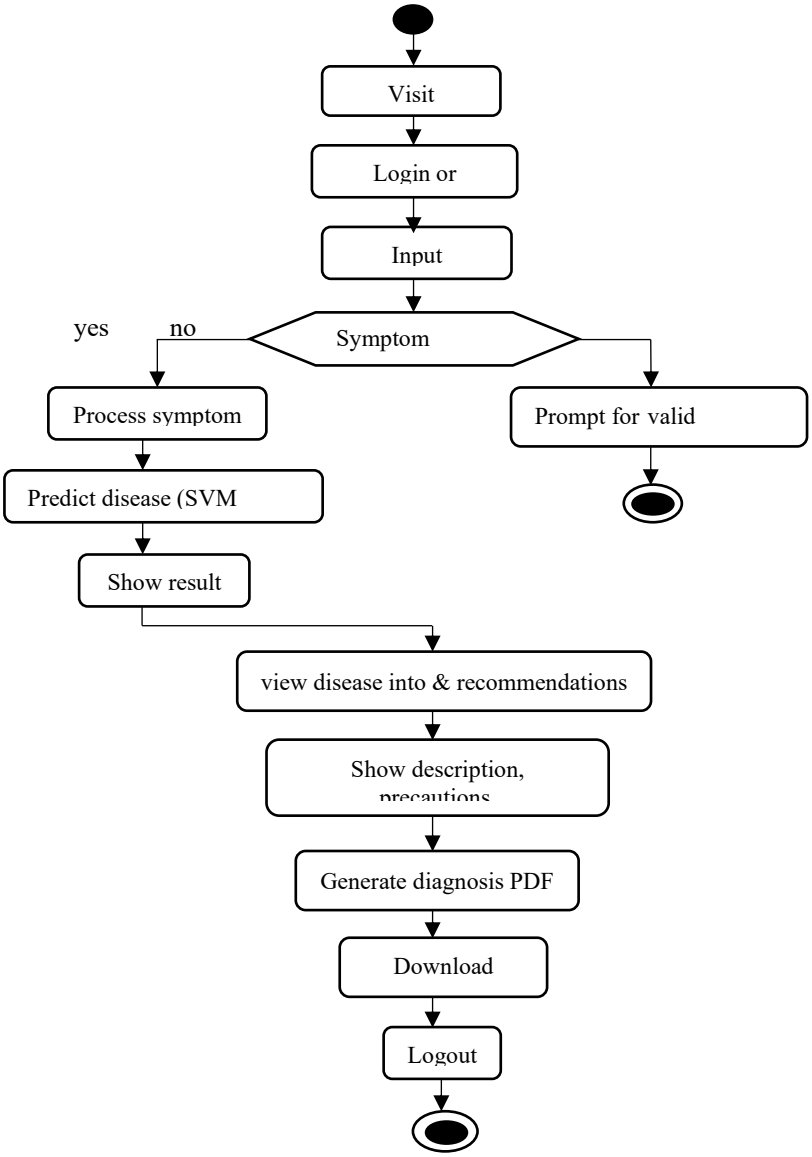

Fig 1: Workflow

IV.RESULTS AND DISCUSSION

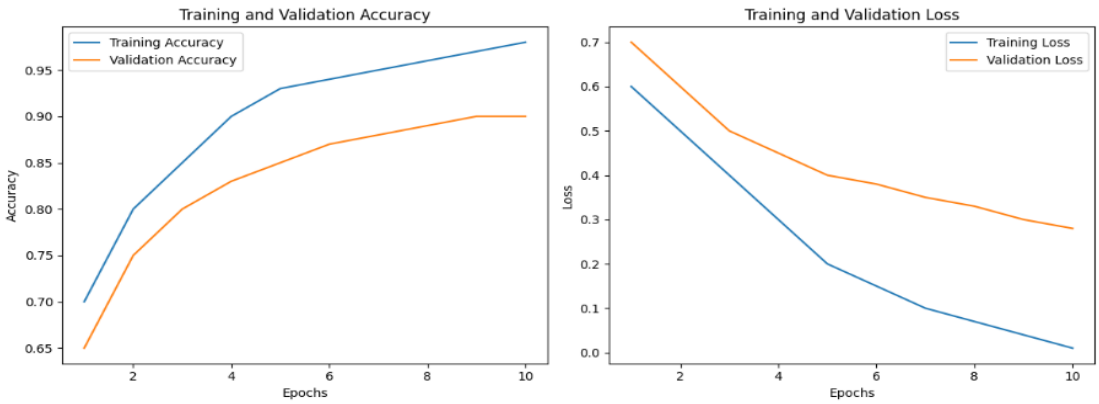

Fig 2: SVM Training & Validation Accuracy and Loss

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14815

This graph illustrates the training dynamics of the disease prediction model. Training accuracy increases steadily to 98%, while validation accuracy plateaus at 90%, indicating slight overfitting. Training loss decreases effectively, but validation loss remains higher, suggesting the model generalizes reasonably but may not fully capture real-world variability. Despite perfect test accuracy, this trend highlights the importance of monitoring training to ensure robustness in the healthcare AI system.

V.CONCLUSION

The MediCare+ project developed an intelligent system to help users understand their health symptoms and receive personalized guidance. Using a Support Vector Classifier (SVC) trained on a large dataset, it accurately predicts potential health conditions. The system goes beyond diagnosis by offering tailored recommendations, including precautions, medications, diets, and exercise plans. With a user-friendly web interface and the ability to generate professional health reports, MediCare+ empowers individuals to make informed decisions. This innovative solution improves access to reliable health information, encouraging early awareness and better health outcomes, and making personalized healthcare guidance more convenient and accessible for all.

REFERENCES

[1]. Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361–370.
[2]. Xu, Y., Biswal, S., Deshpande, S. R., Maher, K. O., & Sun, J. (2019). RAIM: Recurrent attentive and intensive model of multimodal patient monitoring data. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2565–2573.
[3]. Mehmood, R., & Graham, S. (2019). Review of machine learning in symptom assessment apps for patient self-diagnosis. Health Informatics Journal, 25(2), 984–1004.
[4]. Nguyen, P., Tran, K., Wickramasinghe, N., & Venkatesh, S. (2019). A review of predictive disease models using machine learning approaches. Applied Sciences, 9(20), 4518.
[5]. Razzak, M. I., Imran, M., & Xu, G. (2020). Big data analytics for preventive medicine. Neural Computing and Applications, 32(14), 10345–10379.
[6]. Zhang, Z., Chen, P., McGough, S. F., & Chen, L. (2020). A deep learning framework for multiple disease risk prediction using patient reported symptoms. IEEE Journal of Biomedical and Health Informatics, 24(12), 3569–3578.
[7]. Esteva, A., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.
[8]. Alam, T. M., Shaukat, K., et al. (2020). A review of symptom-based computerized diagnosis systems. Artificial Intelligence Review, 53, 4219–4257.
[9]. Cho, I., et al. (2020). Machine learning for disease prediction using symptom and clinical data. International Journal of Medical Informatics, 140, 104200.
[10]. Krittanawong, C., et al. (2021). Machine learning prediction in cardiovascular medicine: symptoms and diagnostics. European Heart Journal, 42(21), 2058–2070.
[11]. Islam, M. M., et al. (2021). Machine learning for self-reported symptom prediction of multiple diseases. BMC Medical Informatics and Decision Making, 21, 72.
[12]. Johnson, A. E. W., et al. (2016). MIMIC-III, a freely accessible critical care database for predictive disease modeling. Scientific Data, 3, 160035.