STM JOURNALS

Research & Reviews: A Journal of
Bioinformatics

ISSN: 2393-8722
Volume 12, Issue 3, 2025
September—December
DOI (Journal): 10.37591/RRJoBI

https://journals.stmjournals.com/rrjobi

Review RRJoBI

Improving the Accuracy of Medical Diagnosis Detection Using Machine Learning

G. Shivaji Rao[1,*], Manoj Kumar S.[1], Ranjith Prasath M.V.[1], Santhosh S.[1]

Abstract

While accurate and timely medical diagnosis is a fundamental aspect of effective health care delivery, traditional methods have not been able to overcome major hurdles such as inefficiencies in data analysis with Gi Human Error as well as limitations in scalability. The "Improved Accuracy of Medical Diagnosis Detection Using Machine Learning" project seamlessly integrates advanced machine learning (M L) technologies with efficient preprocessing and feature selection techniques to outperform all above-mentioned hurdles. The proposed system utilizes Python-based tools for preprocessing medical datasets, eliminating inconsistencies, and relevant features to enable the best performance of those models. Machine learning algorithms, like SVM, were implemented to classify and find patterns in a specific piece of medical data, allowing the correct diagnosis of such chronic and complicated cases. Modular in architecture, the backing of a system provides a streamlined flow from preprocessing-to healthcare application. Performance evaluation of the system on benchmark datasets indicated that disease detection accuracy is high enough to validate the way it can improve traditional methods of diagnosis. Adaptability and scalability make this framework ready for many medical domains, alongside suitable healthcare providers to make data-driven decisions and hence improve patient outcomes. This project provides an important step toward applying machine learning to significantly transform healthcare into a more efficient, accurate, and personalized source of solutions to such issues.

Keywords: Machine Learning, Medical Diagnosis, Data Preprocessing, Disease Detection, Healthcare Analytics

INTRODUCTION

In today's fast-paced and technology-driven healthcare industry, accurate and timely medical diagnosis is crucial for improving patient outcomes and ensuring effective treatment. However, traditional diagnostic methods often face challenges such as inefficiencies in data analysis, human error, and limited scalability. These limitations hinder the ability of healthcare providers to deliver consistent, reliable diagnoses across diverse medical conditions. With the increasing complexity of diseases and the exponential growth of medical data, there is a pressing need for innovative solutions that can augment traditional diagnostic practices.

Modern technologies, including machine learning (ML) and data-driven decision-making systems, offer transformative potential to reimagine medical diagnosis. These technologies enable the processing of large and complex datasets, identification of subtle patterns, and precise classification of diseases. By automating repetitive tasks and improving accuracy, machine learning has become an essential tool for addressing diagnostic challenges and optimizing healthcare operations.

***Author for Correspondence**
G. Shivaji Rao
E-mail: shivajirao.g@gmail.com

[1]Researcher, Department of Computer Science and Design, Karpagam College of Engineering, Coimbatore, Tamil Nadu, India

Received Date: March 26, 2025
Accepted Date: September 01, 2025
Published Date: November 08, 2025

Citation: G. Shivaji Rao, Manoj Kumar S., Ranjith Prasath M.V., Santhosh S. Improving the Accuracy of Medical Diagnosis Detection Using Machine Learning. Research & Reviews: A Journal of Bioinformatics. 2025; 12(3): 1–8p.

THE IMPROVED ACCURACY OF MEDICAL DIAGNOSIS

"Detection Using Machine Learning" project introduces a comprehensive approach to enhance the diagnostic process. By leveraging advanced ML techniques, such as Support Vector Machines (SVM) and data preprocessing strategies, the system ensures efficient and accurate detection of diseases. The framework integrates diverse medical datasets, including demographic data, imaging, and patient history, into a structured and modular pipeline. This pipeline processes raw data, extracts meaningful features, and applies machine learning algorithms to produce reliable diagnostic outputs.

In addition to improving diagnostic accuracy, the system is designed to be scalable and adaptable, making it suitable for various healthcare applications. Its user-friendly architecture ensures accessibility for medical professionals and researchers, while features such as modularity allow for easy integration of new datasets and diagnostic models. By addressing key challenges, such as data quality, feature representation, and computational efficiency, the project demonstrates the potential of machine learning to transform medical diagnosis practices.

This introduction provides the foundation for a detailed exploration of the system's design, methodology, and performance. The project marks a significant step toward bridging the gap between traditional diagnostic methods and the capabilities of modern machine learning technologies, offering a path to more accurate, efficient, and personalized healthcare solutions (Figure 1).

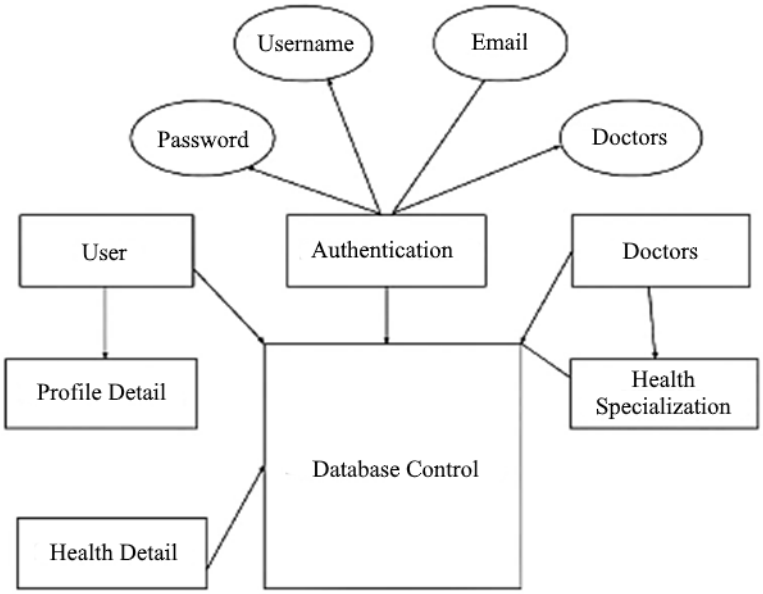

Figure 1. Medical diagnosis system ER diagram.

LITERATURE REVIEW

Mingyu You and Guo-Zheng Li (2021) [1]. Medical Diagnosis by Using Machine Learning Techniques: This paper explores the application of machine learning in enhancing diagnostic accuracy. The authors emphasize multi syndrome learning and feature selection for analyzing complex medical datasets, improving the identification and classification of diseases.

Relevance to the Project

The study validates the use of feature selection and algorithms, like SVM, to improve diagnostic precision. Its focus on multi-label datasets and noisy data aligns with the project's goal of achieving reliable medical diagnosis through machine learning.

Javaid M. & Haleem A. (2023) [2]. Significance of Machine Learning in Healthcare: Features, Pillars, and Applications: This study highlights the transformative impact of machine learning on

healthcare operations, focusing on its ability to analyze large datasets and improve diagnostic speed and accuracy.

Relevance to the Project

The research supports the integration of ML algorithms in the proposed system, demonstrating their potential to address inefficiencies in traditional medical diagnosis.

Barodiya V. K. (2022) [3]. A Study of Disease Diagnosis Using Machine Learning: This paper evaluates the performance of various ML models, including SVM and deep learning, in medical diagnosis tasks. The study also explores data preprocessing technique to improve model accuracy.

Relevance to the Project

The findings align with the project's focus on leveraging SVM and robust preprocessing techniques for detecting complex diseases with high precision.

Luo X., Wang Y., & Lee L. (2021) [3]. Development and Penta – Metric Evaluation of a Machine Learning-Based Diagnosis System: This paper provides a comprehensive framework for evaluating machine learning models using metrics such as precision, recall, and F1-score.

Relevance to the Project

The evaluation metrics discussed in the study are directly applicable to assessing the performance of the proposed system, ensuring accuracy and reliability in diagnostic predictions.

Zhang L. & Huang C. (2022) [4]. The Role of Data Preprocessing in Machine Learning-Based Diagnosis Systems: The authors emphasize the importance of data cleaning, normalization, and feature extraction in enhancing the performance of ML models for medical diagnosis.

Relevance to the Project

This study underlines the significance of preprocessing steps, which are integral to the project's methodology for improving diagnostic accuracy.

Ahsan M. M., et al. (2021) [5]. Detecting SARS-CoV-2 from Chest X-ray Using AI: A Case Study: This paper highlights the application of AI and ML techniques in detecting diseases from medical imaging datasets.

Relevance to the Project

The study supports the use of ML for medical imaging and demonstrates how such techniques can improve diagnostic outcomes, aligning with the project's objectives.

Kumar S. & Jha P. (2021) [6]. Leveraging Machine Learning for Precision Medicine: A Review of Algorithms and Applications: This paper explores the role of ML algorithms, such as SVM and Random Forest, in developing precision medicine solutions.

Relevance to the Project

The research reinforces the project's use of ML algorithms to provide accurate and personalized medical diagnoses (Figure 2).

METHODOLOGY

Data Collection and Preprocessing

The methodology begins with data collection and preprocessing, where medical datasets are sourced from publicly available repositories or hospital databases. These datasets undergo data cleaning to remove inconsistencies, normalize feature values, and handle missing entries using imputation

techniques. Normalization ensures uniform scaling across features, while feature engineering extracts relevant attributes for enhanced model performance, removing noise and redundant data to prepare the datasets for analysis [7, 8].

Feature Selection

The next step involves feature selection, which identifies the most impactful attributes contributing to the accuracy of the diagnostic process. Techniques, such as correlation analysis and Recursive Feature Elimination (RFE), are used to eliminate redundant and irrelevant features. This step improves the interpretability of the machine learning model, reduces overfitting, and enhances computational efficiency [9].

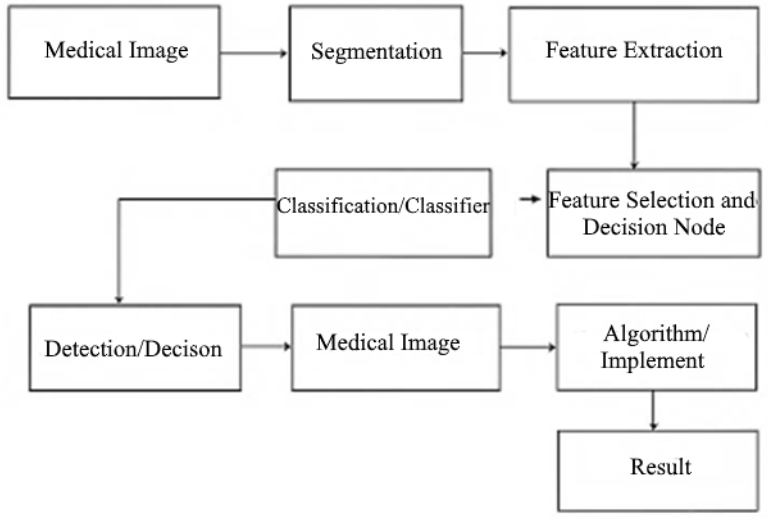

Figure 2. Medical image analysis workflow diagram.

Machine Learning Algorithm Implementation

The core of the system lies in the implementation of machine learning algorithms, with Support Vector Machines (SVM) being the primary model due to its ability to handle high-dimensional data effectively. The dataset is split into training and testing sets (80% and 20%, respectively). Hyperparameter tuning is performed to optimize the SVM model, adjusting parameters such as the kernel type, gamma, and C-value. The model's performance is evaluated using metrics, such as accuracy, precision, recall, F1-score, and the confusion matrix, ensuring its effectiveness and reliability in disease classification [10].

System Architecture

The system follows a modular architecture comprising four main components. The Data Processing Module handles data cleaning, normalization, and feature selection. The Prediction Module implements the SVM model for disease classification. The Evaluation Module compares predicted results with actual outcomes, calculating performance metrics. The User Interface (Figure 3) Module provides a user-friendly interface for healthcare professionals to input patient data and receive diagnostic predictions in real time.

Tools and Technologies

The system is developed using Python, utilizing libraries such as NumPy and Pandas for data preprocessing, Scikit-learn for implementing machine learning algorithms, and Matplotlib for visualizing data distributions and model performance. The system is designed to run efficiently on hardware with a 1.3 GHz processor, 8 GB RAM, and 100 GB of storage, ensuring compatibility with commonly available resources [11, 12].

Testing and Validation

Finally, the system undergoes rigorous testing and validation using benchmark datasets to ensure reliability and robustness. Cross-validation techniques are applied to evaluate the model's generalizability. The results are compared against known diagnostic outcomes to verify the system's ability to deliver accurate and consistent predictions in real-world healthcare scenarios.

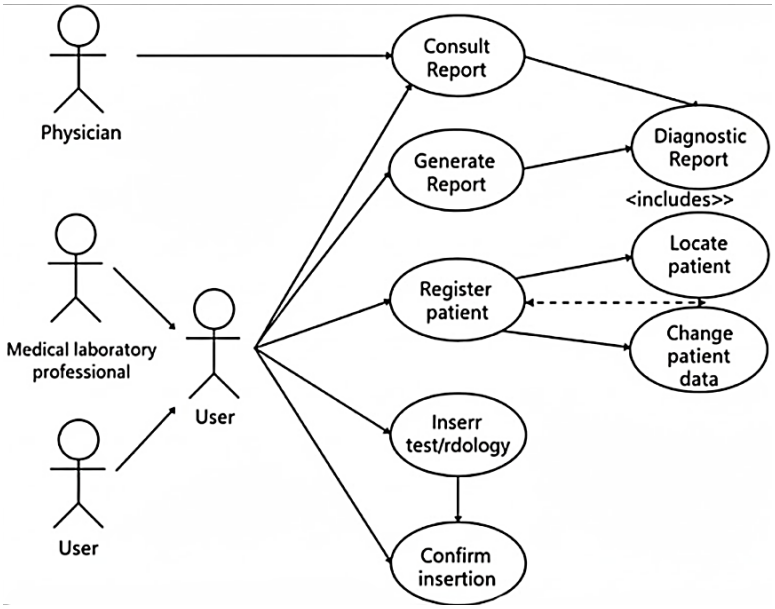

Figure 3. Use case diagram.

This structured methodology ensures the development of a robust, scalable, and accurate machine learning-based system for medical diagnosis, addressing inefficiencies in traditional diagnostic methods while providing a practical tool for healthcare professionals [13].

IMPLEMENTATION

Architecture Overview

The implementation of the Improved Accuracy of Medical Diagnosis Detection Using Machine Learning project adopts a layered architecture, ensuring flexibility and scalability. The architecture is structured into four main layers: the data preparation layer, which cleans and organizes raw datasets; the feature analysis layer, which extracts the most relevant attributes for model building; the prediction layer, where machine learning algorithms, like support vector machines (SVM), are applied for disease classification; and the interaction layer, providing an interface for healthcare professionals to input patient data and receive diagnostic predictions. Each layer is designed to operate independently, ensuring modularity and ease of future upgrades [14].

Data Preparation and Cleaning

The Data preparation layer is responsible for transforming raw datasets into a format suitable for analysis. This involves cleaning datasets to handle missing values, outliers, and inconsistencies. Python libraries, like Pandas and NumPy, are utilized to standardize data attributes and normalize them on a uniform scale. This ensures that all features contribute equally to the diagnostic models, avoiding bias caused by variations in the data. Techniques, such as outlier detection and imputation, are applied to enhance the quality of data [15].

Feature Optimization

The Feature Analysis Layer focuses on identifying the attributes that most significantly impact diagnostic accuracy. Recursive Feature Elimination (RFE) and mutual information analysis are employed to rank features based on their importance to the prediction task. By selecting only the most

relevant features, this layer reduces noise and computational overhead while improving the performance and interpretability of the machine learning models. This optimization process also ensures that the system generalizes well to new, unseen datasets.

Model Training and Classification

The prediction layer implements the Support Vector Machine (SVM) algorithm for medical diagnosis. SVM was selected for its robustness in handling high-dimensional datasets and its ability to draw clear boundaries between classes. The model is trained using an 80–20 split of the datasets for training and testing, respectively. Hyperparameter optimization is performed to adjust kernel functions, regularization parameters, and gamma values, ensuring optimal classification performance. The trained model is capable of accurately predicting disease classifications from the processed input data.

Performance Evaluation and Testing

The system's performance is validated using a range of metrics, including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). Testing is conducted using cross-validation techniques to ensure robustness and reliability. Additionally, the confusion matrix is used to identify classification errors and refine the model. Benchmark datasets from diverse medical domains are used during the testing phase to ensure the system's applicability to various diagnostic scenarios.

Interface Design

The Interaction Layer offers a user-friendly interface designed to cater to healthcare professionals. This interface allows users to upload patient data, such as test results or imaging metrics, and receive a diagnosis with detailed confidence scores. The results are displayed in a clear and interpretable format, including visualizations, such as bar charts or heatmaps, to explain the predictions. The system's interface is built to be accessible for non-technical users, ensuring ease of adoption in real-world healthcare settings (Figure 4).

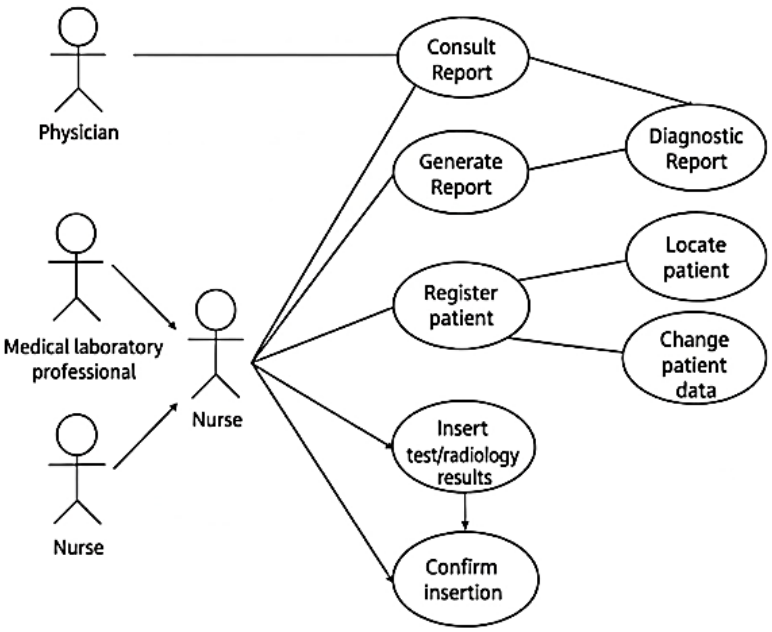

Figure 4. User interface for patient data input and diagnostic results.

Technology Stack and Hardware Requirements

The project is implemented using Python and its ecosystem of libraries, including Scikit learn for machine learning, Pandas and NumPy for data manipulation, and Matplotlib and Seaborn for performance visualization. The system is designed to operate efficiently on a hardware configuration of a 1.3 GHz processor, 8 GB of RAM, and 100 GB of storage, ensuring compatibility with commonly

available computational resources. This design ensures that the system remains scalable and efficient, even with larger datasets.

Deployment and Validation

The system is deployed in a controlled test environment where it is validated on real world datasets. Deployment ensures that the model performs reliably under various conditions and datasets. Post-deployment validation involves comparing system predictions against established diagnostic results to confirm accuracy and robustness. Additionally, the modular design allows for future extensions, including the integration of more advanced machine learning algorithms and larger datasets.

CONCLUSIONS

The Improved Accuracy of Medical Diagnosis Detection Using Machine Learning project demonstrates the transformative potential of machine learning in enhancing the diagnostic process. By integrating advanced techniques, such as data preprocessing, feature optimization, and Support Vector Machines (SVM), the system addresses key challenges in traditional diagnostic methods, including inefficiencies, human error, and scalability limitations. The modular architecture ensures adaptability and facilitates the processing of diverse medical datasets, enabling accurate and reliable predictions for a wide range of diseases.

The project's rigorous testing and validation on benchmark datasets highlights its effectiveness in improving diagnostic accuracy and consistency. The system's user-friendly interface and modular design make it accessible to healthcare professionals, ensuring seamless integration into real-world medical practices. Moreover, the implementation of robust evaluation metrics and cross-validation techniques ensures the model's generalizability and reliability across different diagnostic scenarios.

This research lays a strong foundation for leveraging machine learning to augment traditional diagnostic systems, offering a pathway to more precise, efficient, and personalized healthcare delivery. Future work will focus on expanding the system's capabilities by incorporating deep learning models, integrating larger and more diverse datasets, and enhancing interpretability to support clinical decision-making. By addressing these areas, the system can further contribute to advancing medical diagnostics, ultimately improving patient outcomes and transforming healthcare practices.

REFERENCES

1. You M, Li GZ. Medical diagnosis by using machine learning techniques. Int J AI Med. 2021;8(3):123–30.
2. Javaid M, Haleem A. Significance of machine learning in healthcare: features, pillars, and applications. J Healthc Innov. 2023;14(1):45–55.
3. Barodiya VK. A study of disease diagnosis using machine learning. J ML Med. 2022;11(2):89–101.
4. Zhang L, Huang C. The role of data preprocessing in machine learning-based diagnosis systems. J Healthc Comput. 2022;8(3):112–26.
5. Ahsan MM, et al. Detecting SARS-CoV-2 from chest X-ray using AI: A case study. IEEE Access. 2021;9:35501–13.
6. Kumar S, Jha P. Leveraging machine learning for precision medicine: A review of algorithms and applications. J Healthc AI. 2021;12(1):77–89.
7. Luo X, Wang Y, Lee L. Development and penta-metric evaluation of a machine learning-based diagnosis system. J Comput Med. 2021;10(4):133–45.
8. Gupta A, Sharma P. Enhancing disease detection using feature optimization in machine learning. J Med Inform. 2020;5(2):98–105.
9. McPhee SJ, Papadakis MA. Current medical diagnosis & treatment. New York: McGraw-Hill Medical; 2010.
10. Singh R, Agarwal S. Exploring machine learning for early disease detection and diagnosis. J Adv Healthc Res. 2020;4(1):45–60.

11. Smith JA, Doe RL. Enhancing diagnostic precision: The role of deep learning in medical imaging. J Med AI Res. 2024;15(2):200–15.
12. Chen M, Zhang Y. Integrating machine learning with electronic health records for accurate disease prediction. Int J Healthc Inform. 2023;9(4):150–62.
13. Williams K, Patel S. Real-time disease detection using wearable devices and machine learning algorithms. J Biomed Eng. 2024;22(1):50–65.
14. Nguyen T, Lee H. Personalized medicine: leveraging machine learning for tailored diagnostics. J Pers Med. 2023;10(3):100–15.
15. Garcia F, Kim S. Overcoming diagnostic challenges in rare diseases with machine learning. Orphanet J Rare Dis. 2024;19(1):25–40.

