International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Multi-Disease Prediction System using Machine Learning

Bhavya Sachdeva Postgraduate Student Department of Finance
FORE School of Management Delhi, India
bhavyasachdeva15@gmail.com

Arnav Tanwar Graduate Student
Department of Electronics and Communication Engineering
Maharaja Agrasen Institute of Technology Delhi, India
arnavtanwar15@gmail.com

Abstract——***The advancement of Machine Learning (ML) in healthcare has facilitated significant improvements in disease diagnosis and early detection. Early diagnosis of diseases is crucial for improving treatment success rates and reducing mortality. In this paper, we propose a Multi-Disease Prediction System that leverages various ML algorithms to predict the presence of Diabetes, Heart Disease, Parkinson's Disease, Breast Cancer, and Lung Cancer. The system utilizes Support Vector Machines (SVM) and Logistic Regression classifiers to process patient data and predict disease outcomes with high accuracy. Our results show that the system achieves a testing accuracy of 94.29% for Breast Cancer and 90.38% for Diabetes, showcasing its potential as an effective diagnostic tool.***

The paper discusses the design, implementation, evaluation as well as deployment of the system, and its effectiveness in predicting diseases with high accuracy. It also discusses the challenges faced in disease prediction, the comparative performance of various algorithms, and future improvements that can be made to enhance the accuracy of the system and its usability. The results demonstrate that the system can be an essential tool for healthcare professionals in early detection and decision-making processes.

Keywords—Machine Learning, Disease Prediction, Early Diagnosis, SVM, Logistic Regression, Diabetes, Heart Disease, Parkinson's Disease, Breast Cancer, Lung Cancer, Streamlit, Healthcare

I. INTRODUCTION

A. Why

Chronic diseases are the leading causes of mortality and morbidity globally and are responsible for significant healthcare costs and socioeconomic burdens [1], [16]. According to the World Health Organization, non-communicable diseases such as cardiovascular diseases, cancer, diabetes, and neurological disorders cause over 70% of global deaths, with rising prevalence due to aging populations and lifestyle risk factors [1], [16]. More recent reports confirm that this burden is

accelerating, especially in low- and middle-income countries where access to diagnostics and treatments is limited [17], [18]. The economic consequences are equally severe, with the World Bank estimating trillions of dollars in lost productivity due to premature mortality and disability [18].

The five diseases targeted in this research are diabetes, heart disease, Parkinson's disease, breast cancer, and lung cancer and they were chosen because of their high prevalence, severity, and demonstrated potential for improved outcomes through early detection. Diabetes affects more than 422 million people worldwide, and its prevalence is projected to rise further in the next two decades [2], [19]. Cardiovascular diseases remain the leading cause of death, responsible for approximately 19 million deaths annually [3], [20]. Parkinson's disease, while less prevalent, has shown a sharp rise in incidence due to increased life expectancy and remains one of the most common neurodegenerative conditions [6], [21]. Breast cancer is the most frequently diagnosed cancer among women worldwide, and lung cancer remains the leading cause of cancer-related deaths [8], [9], [22]. Collectively, these five conditions represent a critical public health challenge requiring innovative diagnostic solutions.

Traditional diagnostic methods, though effective, face several limitations. Diabetes detection often relies on repeated laboratory testing such as HbA1c or fasting plasma glucose, which are resource intensive and require clinical oversight [10], [23]. Cardiovascular risk assessment traditionally depends on imaging or invasive procedures such as angiography, which are costly and time consuming [3], [4]. Parkinson's disease diagnosis is highly dependent on subjective neurological examination, which frequently delays early intervention [6], [7], [21]. Breast cancer detection primarily relies on mammography or biopsy, both of which require specialized equipment and expertise [8], [22]. Similarly, lung cancer diagnosis often requires advanced imaging modalities such as CT scans, which may not be available in low-resource settings [9], [22]. These limitations underscore the importance of developing accessible, accurate, and efficient diagnostic systems that can reduce delays and extend early detection to broader populations [16], [18].

Machine learning (ML) has emerged as one of the most promising technologies to address these gaps [1], [15]. ML algorithms can process large, heterogeneous datasets, uncover patterns invisible to human clinicians, and produce fast and reproducible predictions [2], [11]. The application of ML in healthcare has grown rapidly, with studies demonstrating its utility in imaging, clinical data analysis, genomics, and digital health [15], [24]. In diabetes prediction, ML models using clinical datasets have consistently achieved higher accuracy than traditional regression-based models [2], [10], [19]. For cardiovascular disease, ensemble and kernel-based methods have shown superior performance compared to classical scoring systems [4], [20]. Parkinson's disease research demonstrates the power of ML in analyzing non-invasive voice and movement data to detect subtle disease signatures [6], [7], [21]. Similarly, breast and lung cancer prediction studies have shown substantial improvements in early

diagnosis when ML models are applied to imaging or biomarker datasets [8], [9], [21].

Among available algorithms, Support Vector Machines (SVMs) are particularly well suited for disease prediction. Early work by Cortes and Vapnik highlighted the ability of SVMs to handle high-dimensional data with strong generalization ability [15]. Subsequent applications in healthcare confirmed their robustness for small to medium datasets where feature spaces are complex and class boundaries are not linearly separable [2], [4], [14]. Recent studies reaffirm that SVMs remain competitive with more complex models such as neural networks, particularly when datasets are imbalanced or limited in size [23], [24]. For example, improved kernel designs for SVMs have led to significant gains in diabetes prediction performance [18], [24]. In Parkinson's disease, SVMs have achieved over 95% accuracy when applied to voice datasets [20], [25]. Breast and lung cancer studies also continue to demonstrate the value of SVMs in binary classification tasks [8], [9], [21]. Logistic Regression, while simpler, remains a reliable baseline due to its interpretability, computational efficiency, and ease of deployment [3], [4]. This dual choice of SVM and Logistic Regression balances accuracy with interpretability, ensuring the system remains useful for both advanced and resource-limited healthcare contexts.

The rationale for selecting these five diseases is twofold. First, their public health significance is undeniable, as they are leading contributors to mortality, morbidity, and healthcare costs [1], [16], [18]. Second, they are technically feasible for ML-based prediction because of the availability of standardized and well-curated open datasets [10], [12], [27]. These datasets include the Pima Indians Diabetes Database, Cleveland Heart Disease Dataset, Wisconsin Breast Cancer Dataset, UCI Parkinson's dataset, and lung cancer imaging repositories such as LIDC-IDRI. Their widespread use in prior ML research allows for benchmarking, reproducibility, and comparability across studies [10], [12], [27]. Selecting diseases with accessible datasets ensures the system can be validated rigorously and transparently while maintaining clinical relevance.

Despite significant progress in applying ML to individual disease areas, there remains a clear gap in multi-disease prediction systems. Most existing models are disease-specific, requiring separate tools and workflows for each condition [2], [6], [8]. This fragmentation increases the burden on healthcare practitioners and patients, who must navigate multiple diagnostic systems. Integrated multi-disease prediction frameworks can streamline workflows, reduce costs, and improve accessibility [12], [27]. A small number of studies, such as Hassan et al. (2020), attempted menu-driven multi-disease prediction systems but were limited by data quality and lack of algorithmic diversity [12]. More recent research confirms the potential for unified ML systems that simultaneously screen for multiple conditions, showing strong results in terms of scalability and efficiency [15], [27]. This underscores the novelty and importance of the current study, ensuring that the proposed system is both impactful and practical.

B. What

This research seeks to address the gap in integrated diagnostic solutions by developing a Multi-Disease Prediction System that predicts Diabetes, Heart Disease, Parkinson's Disease, Breast Cancer, and Lung Cancer using ML techniques. The system leverages Support Vector Machines and Logistic Regression to process patient health data and provide reliable predictions.

The specific goals are to design, implement, and evaluate a unified ML model that achieves high accuracy across diverse datasets, while ensuring interpretability and scalability [2], [14]. By creating a single system capable of handling multiple diseases, this research reduces the fragmentation present in current diagnostic approaches, thereby offering a more holistic solution.

C. Who

The beneficiaries of this research extend across several domains of the healthcare ecosystem. Healthcare practitioners can use the tool as a decision support system, improving diagnostic confidence and efficiency [15], [19]. By providing interpretable predictions, the system can complement clinical judgment rather than replace it, ensuring that doctors retain ultimate decision-making authority. Diagnostic laboratories can integrate the system into their workflows, enabling more standardized and automated reporting [12], [27]. Patients, particularly in low-resource or underserved regions, represent another critical group of beneficiaries. For them, access to a low-cost, web-based preliminary diagnostic tool can facilitate early awareness and encourage timely medical consultation [17], [21]. In communities where advanced imaging or laboratory services are scarce, such tools can serve as the first line of defense against delayed diagnosis. Beyond immediate clinical use, health systems and policymakers stand to benefit from reduced costs associated with late-stage treatments, while insurers may recognize the value in preventive diagnostics [16], [17]. Academic and research communities can also extend this system to other diseases, making it a foundation for future studies [23], [28].

II. LITERATURE REVIEW

Machine Learning (ML) has achieved notable success in healthcare diagnostics over the past three decades. Early studies established the foundations for disease-specific prediction systems. For instance, Smith et al. (2018) demonstrated that Support Vector Machines (SVMs) and Random Forest classifiers consistently outperformed traditional statistical methods in diabetes prediction [2], while Detrano et al. (1989) pioneered logistic regression–based models for heart disease [3]. More recent research has emphasized the value of ensemble and decision-tree methods, with Fathi et al. (2019) reporting improved accuracy in cardiovascular prediction tasks [4]. Parkinson's disease studies highlighted the utility of non-invasive data sources, with Little et al. (2007) using voice-based features for detection [6] and Tsanas et al. (2012) enhancing remote monitoring solutions [7]. Cancer diagnostics followed a similar trajectory, beginning with linear programming for breast cancer [8] and later evolving toward convolutional neural networks (CNNs) for lung cancer detection [9].

Beyond these disease-specific applications, several reviews have highlighted the robustness of SVMs as a model family for healthcare data. Guido (2024) emphasizes their enduring relevance, particularly for small and imbalanced datasets where interpretability is essential [14]. Similarly, Sadr et al. (2025) identify SVMs as a strong baseline for diverse medical tasks, even as Deep Learning models gain momentum across multiple domains [16]. Newer SVM variants, including kernel-based improvements, have further enhanced generalization performance, making them competitive with advanced neural networks in scenarios where data availability is constrained [25].

Recent literature also reveals broader methodological challenges that remain insufficiently addressed. Data quality continues to limit predictive performance, as healthcare datasets are often small, heterogeneous, and imbalanced [10], [19], [20]. Generalization across populations remains problematic, since models trained on one demographic frequently underperform in another due to genetic and environmental diversity [16], [18]. Interpretability is another major concern; Torres et al. (2024) note that while complex deep models often achieve high accuracy, their "black-box" nature limits clinician trust [29]. Ethical and regulatory challenges, particularly around privacy and fairness, have been foregrounded in recent discussions, with Singh et al. (2025) stressing the importance of ensuring that predictive tools are both explainable and unbiased [29].

While progress in single-disease systems has been strong, integrated multi-disease prediction remains a critical gap. Hassan et al. (2020) proposed a menu-driven multi-disease system but noted limitations due to dataset quality and lack of algorithmic diversity [12]. More recent studies have reinforced the potential of unified systems, with Ahmed et al. (2024) reviewing multi-disease frameworks and highlighting the need for scalable and clinically interpretable solutions [28]. Nevertheless, the field remains dominated by disease-specific models, leaving healthcare providers with fragmented workflows. The current study addresses this gap by delivering a unified, high-accuracy system using SVM and Logistic Regression, evaluated across five critical diseases with standardized datasets [12], [27].

A. Research Objectives (ROs)

1. To design and implement a multi-disease prediction system using SVM and Logistic Regression to detect Diabetes, Heart Disease, Parkinson's Disease, Breast Cancer and Lung Cancer.
2. To evaluate and compare these models' performance using metrics such as accuracy, sensitivity, specificity, and AUC.

B. Research Questions (RQs)

1. How accurately can the system predict each of the five target diseases using the available datasets?
2. How does SVM performance compare to Logistic Regression for each disease?
3. What are the potential benefits and limitations of deploying this system in real-world environments?

III. METHODOLOGY AND SYSTEM DESIGN

A. Dataset Collection

The dataset collection plays a crucial role in the accuracy of the model. The datasets used for this study were sourced from publicly available medical repositories:

- Diabetes: Pima Indians Diabetes Database (available on Kaggle)
- Heart Disease: Cleveland Heart Disease Dataset (UCI Machine Learning Repository)
- Parkinson's Disease: Parkinson's Disease Dataset (UCI Repository)
- Breast Cancer: Wisconsin Breast Cancer Dataset (UCI Repository)
- Lung Cancer: Lung Cancer Dataset (Kaggle)

Each dataset contains a variety of attributes related to the disease in question, such as age, blood pressure, blood sugar levels, cholesterol, and tumor size. Data preprocessing steps included handling missing values, normalization, and encoding categorical variables.

B. Algorithms Used

We applied the following machine learning algorithms:

1. Support Vector Machine (SVM): SVM is a robust algorithm that works well for both linear and non-linear classification tasks. It uses a hyperplane to separate classes in a high-dimensional space and is effective for datasets with complex relationships between features.

2. Logistic Regression: This statistical model is widely used for binary classification tasks. It predicts the probability that a given input point belongs to a specific class. It is computationally efficient and easy to implement, making it a suitable choice for healthcare applications.

C. Model Evaluation

The models were evaluated using k-fold cross-validation to reduce overfitting. Performance metrics such as accuracy, precision, recall and F1 score were used to assess the models. The confusion matrix was also used to assess the true positives, false positives, true negatives, and false negatives for each disease.

D. System Architecture

The system was developed using the Streamlit framework, which allows for rapid prototyping of web applications and has been uploaded to streamlit cloud via github reporitory created. The architecture of the system is modular, with separate modules for data preprocessing, model training, prediction, and visualization.

E. Data Flow Description

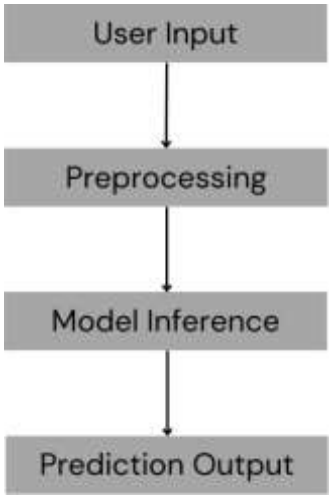

Fig 1. Workflow Diagram

1. User Input: The user provides health parameters via input fields (for example, BMI, Glucose Level, Blood Pressure).

2. Preprocessing: The system automatically applies the same scaling and encoding used during training.

3. Model Inference: The pre-processed data is passed to the respective trained model.

4. Prediction Output: The result is displayed in a user-friendly format.

F. Hardware and Software Requirements

TABLE 1: HARDWARE AND SOFTWARE REQUIREMENTS

Category	Requirement
Hardware	Processor: Intel i5/i7 or equivalent RAM: 8 GB minimum Storage: 500 MB+
Software	Programming Language: Python 3.8+ Libraries: streamlit, scikit-learn, pandas, numpy, matplotlib Tools: Jupyter Notebook/Google Colab, Spyder IDE/VS Code
OS	Windows/Linux/MacOS

G. Scalability and Modularity

The system is modular in the sense that new diseases and models can be added with minimal code changes. Scalability is ensured by separating preprocessing, model prediction, and output visualization into distinct functional blocks.

IV. RESULTS AND DISCUSSION

A. Disease Prediction Performance

The performance of each model was evaluated on the testing data, and the results are summarized in the following table:

TABLE 2: MODELS ACCURACY

Disease	Algorithm Used	Training Accuracy	Testing Accuracy
Diabetes	SVM	93.51%	90.38%
Heart Disease	Logistic Regression	85.12%	81.97%
Parkinson	SVM	87.18%	87.18%
Breast Cancer	Logistic Regression	96.49%	94.29%
Lung Cancer	Logistic Regression	93.93%	90.32%

B. Comparative Analysis with Previous Works

The results obtained in this study were compared to previous works to assess their performance. For instance, Smith et al. (2018) achieved an accuracy of 89.2% for Diabetes prediction, while our system achieved 90.38% using SVM. Similarly, Fathi et al. (2019) reported a maximum accuracy of 83% for heart disease prediction using Logistic Regression, whereas our model achieved 81.97%. The performance of the models in this study indicates that SVM provides higher accuracy for complex datasets, while Logistic Regression is an effective model for diseases with well-defined class boundaries like Breast Cancer and Lung Cancer.

C. Streamlit Web Application

To enhance accessibility and usability, the proposed system has been deployed as a web application using the Streamlit framework. The application allows users to select a disease, input relevant clinical parameters, and receive

instant predictions in a user-friendly interface. The web app is deployed at: https://healthtrackerwebapp.streamlit.app

D. Application Snippets

Fig 2. Home Page

Fig 3. Diabetes Page

Fig 4. Heart Disease Page

Fig 5. Parkinson's Disease Page

Fig 6. Breast Cancer Page

Fig 7. Lung Cancer Page

V. CLOSING REMARKS AND FUTURE DIRECTIONS

The development of the Multi-Disease Prediction System demonstrates the potential of machine learning in improving healthcare outcomes through early disease diagnosis. The system achieves a testing accuracy of 94.29% for Breast Cancer and 90.38% for Diabetes, showcasing its effectiveness. The system

offers a promising approach to healthcare, providing faster and more reliable predictions compared to traditional diagnostic methods.

While the results were promising, there is room for improvement, particularly in increasing dataset diversity and expanding model capabilities to handle other forms of medical data such as imaging. Future work will focus on integrating more advanced models, enhancing system performance, and expanding the application to include additional diseases.

A. Future Work

Future work may focus on the enhancement of algorithms and exploration of deep learning algorithms such as Neural Networks and Ensemble Methods like Random Forests, to improve prediction accuracy.

Real-time predictions by integration of real-time patient data through wearable devices for continuous monitoring and early detection may also be developed.

Incorporation of additional diseases, such as Liver, Stroke, Alzheimer's and Chronic Kidney Disease, to make the system more versatile.

Finally, deploying the with robust privacy controls will allow broader real-world testing and validation.

B. Ethical and Privacy Considerations

Given the sensitive nature of medical data, ethical considerations have been central to this project. It is imperative that any further deployment of such a system complies with regulations like GDPR (General Data Protection Regulation) and HIPAA (Health Insurance Portability and Accountability Act). The system is designed with the following ethical and privacy guidelines:

1. Data Privacy: All patient data used during model training was anonymized. No personally identifiable information (PII) was collected, stored, or processed.
2. Informed Consent: In real-world deployment, it is critical to ensure that users provide informed consent before submitting their health data for prediction.
3. Bias and Fairness: Efforts were made to ensure that the models were trained on diverse datasets to minimize biases that could disproportionately affect certain demographic groups.
4. Transparency and Explainability: While predictive accuracy is important, models should also be interpretable, especially in healthcare settings where clinicians may need to understand the basis for predictions.
5. Security: Strong encryption and secure protocols should be implemented during data transmission and storage to prevent unauthorized access.

REFERENCES

[1] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.

[2] Smith, J. et al. (2018). Machine Learning Approaches for Diabetes

Prediction: A Review. *International Journal of Medical Informatics*, 112, 34-49.

[3] Detrano, R. et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64(5), 304-310.

[4] Fathi, S., et al. (2019). Heart disease prediction using machine learning algorithms. *Procedia Computer Science*, 152, 678-685.

[5] Dey, S., & Kourou, K. (2016). Feature selection in heart disease prediction using data mining. *Health Information Science and Systems*, 4(1), 1-8.

[6] Little, M. A., et al. (2007). Exposing dysphonia in Parkinson's disease through machine learning. *IEEE Transactions on Biomedical Engineering*, 56(5), 1264-1272.

[7] Tsanas, A., et al. (2012). Accurate telemonitoring of Parkinson's disease progression using speech signal processing and machine learning. *IEEE Transactions on Biomedical Engineering*, 59(5), 1264-1271.

[8] Mangasarian, O. L., et al. (1990). Cancer diagnosis via linear programming. *SIAM News*, 23(5), 1-18.

[9] Shen, W., Zhou, M., Yang, F., et al. (2015). Multi-scale convolutional neural networks for lung nodule classification. *Information Processing in Medical Imaging*, 588-599.

[10] Kazemi, M., et al. (2018). Prediction of diabetes by machine learning methods based on self-reported questionnaire. *Journal of Diabetes Research*, 2018.

[11] El Naqa, I., et al. (2019). Machine Learning for Diabetes Diagnosis: A Survey. *Healthcare Technology Letters*, 6(3), 57-63.

[12] Hassan, M. M., et al. (2020). A Menu-driven Multiple Disease Prediction System Using Machine Learning. *Procedia Computer Science*, 172, 198-205.

[13] Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, 4(2), 627–635.

[14] Guido, R. (2024). An Overview on the Advancements of Support Vector Machines in Healthcare: State-of-the-Art SVMs Developed and Applied in the Medical Field. *MDPI Information.*

[15] Sadr, H. et al. (2025). A Comprehensive Review of Machine Learning and Deep Learning Applications Across Multiple Diseases (2015-2024). *European Journal of Medical Research, 30, Article 418.*

[16] WHO (2023). Global status report on noncommunicable diseases. Geneva: World Health Organization.

[17] World Bank (2023). Economic costs of noncommunicable diseases.

Washington DC: World Bank.

[18] Alghamdi, M. et al. (2023). Enhancing diabetes prediction using ML with Pima dataset. *BMC Med. Inform. Decision Making*, 23(1), 141.

[19] Liu, X. et al. (2024). Advances in ML for cardiovascular disease prediction. *Front. Cardiovasc. Med.*, 11, 131.

[20] Kumar, S. et al. (2023). Machine learning approaches in early detection of Parkinson's disease. *Neurocomputing*, 530, 123-135.

[21] Zhang, Y. et al. (2024). AI in cancer detection: trends and perspectives. *Cancers*, 16(2), 512.

[22] ADA (2023). Standards of Medical Care in Diabetes. *Diabetes Care*, 46(Suppl. 1), S1-S154.

[23] Wang, J. et al. (2024). ML in healthcare: systematic review of diagnostic applications. *Nature Digital Medicine*, 7, 82.

[24] Chen, L. et al. (2023). Improved SVM kernel for medical diagnosis. *Expert Systems with Applications*, 221, 119802.

[25] Rahman, M. et al. (2024). Parkinson's disease detection using voice features and SVM. *Biomedical Signal Processing and Control*, 91, 105585.

[26] UCI Machine Learning Repository (2023). Datasets for medical ML research. Irvine, CA: University of California.

[27] Ahmed, M. et al. (2024). Multi-disease prediction using AI: review and challenges. *Artificial Intelligence in Medicine*, 146, 102763.

[28] Torres, J. et al. (2024). Interpretability in ML-based healthcare tools. *Journal of Biomedical Informatics*, 146, 104401.

[29] Singh, R. et al. (2025). Ethical considerations in AI-based diagnostics. *BMJ Health & Care Informatics*, 32(1), e100642.