Machine Learning Approaches for Multiple Disease Prediction a Comprehensive Review

Shridhar Behera[1], Aakanksha Sahu[2], Adarsh Kumar Soni[3]
[1]*Department of Computer Science and Engineering, RSR-Rungta college of engineering and technology, Bhilai (C.G), India*
[2,3] *Guide by Assistant Professor Department of Computer Science and Engineering, RSR-Rungta College of Engineering and Technology, Bhilai (C.G.), India*

***Abstract*—Chronic and non-communicable diseases such as diabetes mellitus, cardiovascular disease, and Parkinson's disease continue to impose a substantial burden on global healthcare systems, making early and accurate diagnosis a pressing clinical priority. Over the past decade, machine learning (ML) has emerged as a powerful tool for building data-driven clinical decision support systems capable of predicting disease risk from structured patient records and biomedical signals. This review synthesizes findings from fifteen peer-reviewed studies published between 2007 and 2025 that apply supervised learning algorithms—including Support Vector Machines, Logistic Regression, Random Forest, K-Nearest Neighbour, Naive Bayes, Decision Trees, ensemble methods, and deep learning models—to disease prediction tasks. The reviewed literature is organized around three disease domains and a fourth, emerging category of unified multi-disease prediction frameworks. A comparative analysis of algorithmic performance, dataset characteristics, and methodological choices is presented in tabular form. The review further identifies recurring research gaps, including limited dataset diversity, weak external validation, inconsistent preprocessing pipelines, and the scarcity of genuinely integrated multi-disease platforms. Finally, promising directions for future research are outlined, spanning federated learning, explainable artificial intelligence, multimodal data fusion, and real-time clinical deployment.**

***Index Terms*—Machine Learning, Disease Prediction, Diabetes, Heart Disease, Parkinson's Disease, Classification Algorithms, Healthcare Informatics, Computer-Aided Diagnosis.**

I. INTRODUCTION

The global health landscape has shifted decisively toward non-communicable diseases (NCDs) as the dominant cause of morbidity and mortality. The World Health Organization estimates that NCDs, including diabetes and cardiovascular conditions, account for the majority of deaths worldwide each year, while neurodegenerative disorders such as Parkinson's disease represent a rapidly growing burden associated with ageing populations. A common thread across these conditions is that early detection substantially improves treatment outcomes and reduces long-term healthcare costs, yet conventional diagnostic pathways often rely on specialist consultation, invasive testing, or symptom progression that delays intervention until the disease has already advanced.

Machine learning offers an alternative diagnostic pathway by learning complex, non-linear relationships between clinical or physiological features and disease outcomes directly from historical patient data. Unlike traditional rule-based clinical scoring systems, ML classifiers can adapt to large, high-dimensional datasets, uncover latent feature interactions, and generate probabilistic risk estimates that support, rather than replace, clinical judgement. Consequently, a substantial body of research has emerged applying algorithms such as Support Vector Machines (SVM), Random Forest, Logistic Regression, K-Nearest Neighbour (KNN), Naive Bayes, Decision Trees, and gradient boosting methods to the prediction of diabetes, heart disease, and Parkinson's disease using well-known public datasets.

Despite this volume of work, the literature remains fragmented. Most studies address a single disease in isolation, employ inconsistent preprocessing and validation protocols, and report accuracy figures that are difficult to compare directly due to differences in dataset partitioning, cross-validation strategy, and

feature engineering choices. Furthermore, relatively few studies attempt to unify multiple disease prediction tasks within a single deployable system, despite the practical appeal of such platforms in resource-constrained or primary-care settings where access to multiple specialists is limited.

Problem Statement: There is a clear need for a consolidated, comparative understanding of how different machine learning techniques perform across diabetes, cardiovascular disease, and Parkinson's disease prediction tasks, and for a systematic identification of the methodological gaps that prevent these research efforts from translating into robust, generalizable, real-world clinical tools. This review addresses that need by synthesizing the methodologies, datasets, and outcomes reported across fifteen representative studies, comparing their techniques in a structured format, and articulating the research gaps and future directions that should guide the next generation of ML-based disease prediction systems.

The remainder of this paper is organized as follows. Section II presents a detailed literature review organized by disease domain. Section III offers a comparative analysis of the techniques discussed, presented in tabular form. Section IV highlights the research gaps identified across the reviewed literature. Section V discusses future research directions, and Section VI concludes the paper.

II. LITERATURE REVIEW

A. Machine Learning for Diabetes Prediction

Diabetes mellitus has been one of the most extensively studied conditions in the ML-based prediction literature, largely owing to the widespread availability of the PIMA Indian Diabetes Dataset. Sisodia and Sisodia [1] compared Decision Tree, SVM, and Naive Bayes classifiers on this dataset and reported that Naive Bayes achieved the highest accuracy among the three, underscoring the suitability of probabilistic classifiers for relatively small, imbalanced clinical datasets. Kavakiotis et al. [2] conducted a broader survey of data mining methods applied to diabetes research, observing that ensemble and hybrid approaches consistently outperformed single classifiers, while also noting that most studies relied on a narrow set of public datasets, limiting generalizability.

Maniruzzaman et al. [3] undertook a comparative evaluation of multiple classification paradigms for diabetes diagnosis and demonstrated that performance varied considerably depending on the feature selection strategy employed prior to classification, suggesting that preprocessing choices can matter as much as algorithm selection. Zou et al. [4] applied Random Forest and J48 decision tree models to a large hospital-based dataset and achieved strong predictive performance, while also showing that the most clinically informative features (such as plasma glucose concentration and body mass index) aligned well with established medical risk factors, lending interpretability to their model outputs. Hasan et al. [5] proposed an ensembling framework that combined several base classifiers through a weighted voting mechanism, reporting improved robustness compared to any individual model. Islam et al. [6] and Qawqzeh et al. [7] further extended this body of work by exploring early-stage risk prediction using mixed categorical-clinical features and logistic regression-based modelling of photoplethysmogram waveform data respectively, illustrating the diversity of feature modalities being explored for diabetes risk stratification.

B. Machine Learning for Heart Disease Prediction

Cardiovascular disease prediction research has similarly converged around the Cleveland Heart Disease dataset, with early contributions such as Detrano et al. [8] establishing probability-based diagnostic algorithms that served as a benchmark for later ML approaches. Alizadehsani et al. [9] applied data mining techniques specifically for coronary artery disease diagnosis, demonstrating that combining multiple weak classifiers improved diagnostic sensitivity relative to single-model baselines. Mohan et al. [10] proposed a hybrid approach that combined Random Forest with a linear model, reporting accuracy improvements attributed to the complementary strengths of tree-based and linear decision boundaries.

Shah et al. [11] benchmarked several standard classifiers, including KNN, Decision Tree, and Naive Bayes, finding that KNN provided the most consistent performance on the Cleveland dataset across varying train-test splits. Bharti et al. [12] extended this line of inquiry by integrating deep learning architectures alongside conventional ML

classifiers, reporting that deep neural networks could match or exceed traditional models given sufficient data, though at the cost of reduced interpretability—a trade-off that recurs throughout the literature. Swathy and Saruladha [13] concluded that, for the relatively small and tabular nature of available cardiac datasets, classical ML models such as SVM and Logistic Regression often remained competitive with, and computationally cheaper than, deep learning alternatives. Peng et al. [14] applied XGBoost with Bayesian hyperparameter optimization, reinforcing that careful model tuning is itself a meaningful contribution to predictive accuracy.

C. Machine Learning for Parkinson's Disease Detection

Parkinson's disease prediction research has taken a markedly different methodological path, centering on acoustic and dysphonia-based biomarkers rather than conventional clinical measurements. Little et al. [15] pioneered the use of nonlinear, fractal-scaling speech features for voice disorder detection, establishing measurement techniques that remain foundational to subsequent Parkinson's prediction studies. A related study by Little et al. [16] evaluated the suitability of dysphonia measurements for remote telemonitoring applications, demonstrating that voice-based biomarkers could plausibly support low-cost, non-invasive disease tracking outside clinical settings.

Shahbakhi et al. [17] combined genetic algorithm-based feature selection with SVM classification, reporting that reducing the dimensionality of voice feature sets before classification improved both accuracy and computational efficiency. Grover et al. [18] applied deep learning techniques to predict not merely the presence but the severity of Parkinson's disease, representing a shift toward more clinically nuanced, ordinal prediction tasks. Ozkaya et al. [19] found that distance-based classifiers such as KNN frequently outperformed kernel-based methods on voice feature spaces, a finding consistent with several multi-disease prediction studies. Tracy et al. [20] explored deep phenotyping methods that combined voice biomarkers with broader behavioral signals, pointing toward the increasing role of multimodal data fusion in neurodegenerative disease research.

D. Multi-Disease and Integrated Prediction Frameworks

A smaller but growing body of literature has begun to address the integration of multiple disease prediction tasks within unified systems. Reddy et al. [21] proposed an integrated multi-disease prediction framework intended for mobile health applications, combining separately trained models for several conditions behind a common interface, and reported that such integration improved system usability without materially compromising individual model accuracy. Yadav and Pal [22] proposed an ensemble-based multi-disease prediction approach for healthcare data, demonstrating that ensembling strategies effective for single-disease prediction could be extended, with modification, to multi-disease contexts. Together, these studies suggest that while unified prediction platforms are technically feasible, the field has not yet converged on standardized architectures, evaluation protocols, or deployment practices for such systems.

III. COMPARATIVE ANALYSIS OF TECHNIQUES

Table I summarizes the algorithmic approaches, target diseases, datasets, and reported performance characteristics of the studies discussed in Section II, providing a consolidated basis for cross-study comparison.

TABLE I Comparative Summary of Machine Learning Techniques for Disease Prediction

Disease Domain	Technique(s)	Dataset	Reported Strength	Reported Limitation	Study
Diabetes	Naive Bayes, SVM, Decision Tree	PIMA Indian Diabetes (768 records)	High accuracy with probabilistic interpretability	Small dataset size limits generalizability	Sisodia & Sisodia [1]
Diabetes	Ensemble voting classifiers	PIMA / hospital records	Improved robustness to noise	Increased computational and tuning overhead	Hasan et al. [5]
Diabetes	Random Forest, J48	Large hospital dataset	Interpretable, clinically	Dataset not publicly benchmarke	Zou et al. [4]

			aligned features	d	
Heart Disease	Hybrid Random Forest + Linear Model	Cleveland Heart Disease (303 records)	Combines tree and linear decision boundaries	Limited external validation	Mohan et al. [10]
Heart Disease	KNN, Decision Tree, Naive Bayes	Cleveland Heart Disease	Consistent performance across splits	Sensitive to feature scaling and k selection	Shah et al. [11]
Heart Disease	Deep Neural Networks + ML hybrid	Cleveland / combined datasets	High accuracy with sufficient data	Reduced interpretability; data-hungry	Bharti et al. [12]
Heart Disease	XGBoost + Bayesian optimization	Cleveland / UCI extended sets	Performance gains via tuning	Risk of overfitting without careful validation	Peng et al. [14]
Parkinson's Disease	SVM + Genetic Algorithm feature selection	Oxford voice dataset (195 recordings)	Improved efficiency via dimensionality reduction	Small sample size; voice-only features	Shahbakhi et al. [17]
Parkinson's Disease	KNN, SVM comparative analysis	Oxford voice dataset	KNN often outperforms kernel methods	Limited demographic diversity in dataset	Ozkaya et al. [19]
Parkinson's Disease	Deep Learning (severity prediction)	Voice + clinical severity scores	Enables ordinal severity prediction	Requires larger labelled severity data	Grover et al. [18]
Multi-Disease	Ensemble framework across diseases	Combined PIMA, Cleveland, Oxford-type sets	Single-interface usability across conditions	Lacks standardized cross-disease evaluation	Reddy et al. [21]
Multi-Disease	Ensemble ML for multi-condition prediction	Aggregated healthcare datasets	Extends single-disease ensembling principles	No unified deployment or validation protocol	Yadav & Pal [22]

IV. RESEARCH GAPS

Synthesis of the reviewed literature reveals several recurring limitations that constrain the clinical translation of ML-based disease prediction research:

- Dataset homogeneity: A large proportion of studies rely on the same small set of public benchmark datasets (PIMA, Cleveland, Oxford Parkinson's), which restricts

demographic diversity and raises concerns about overfitting to dataset-specific idiosyncrasies.

- Inconsistent evaluation protocols: Reported accuracy figures are frequently not directly comparable across studies due to differing train-test splits, cross-validation strategies, and the inconsistent use of metrics beyond raw accuracy.
- Limited external and clinical validation: Very few studies validate trained models on independent, externally collected patient cohorts or in real clinical workflows.
- Interpretability-performance trade-off: Higher-performing ensemble and deep learning approaches often sacrifice the interpretability that clinicians require for trust and adoption.
- Fragmented multi-disease integration: Genuine multi-disease prediction platforms remain rare, amounting mostly to independently trained models bundled behind a shared interface.
- Underexplored multimodal and longitudinal data: Most studies use a single data modality and a single time point, despite the diagnostic value of combining modalities and tracking patients over time.

V. FUTURE SCOPE

Building on the identified gaps, several promising directions can guide future research in machine learning-based disease prediction. First, the development and public release of larger, more demographically diverse datasets—potentially through federated data-sharing consortia—would directly address the dataset homogeneity gap while preserving patient privacy through federated learning techniques that train models across institutions without centralizing sensitive data. Second, greater adoption of explainable AI (XAI) techniques, such as SHAP and LIME, could help reconcile the interpretability-performance trade-off.

Third, future multi-disease systems should move toward jointly optimized architectures that share representations across related conditions, potentially using multi-task learning frameworks that exploit shared risk factors across diabetes, cardiovascular disease, and other comorbid conditions. Fourth, the integration of multimodal data—combining structured clinical records, voice or imaging biomarkers, and wearable sensor streams—represents a natural extension of current single-modality approaches. Finally, prospective clinical validation studies, conducted in collaboration with healthcare institutions and regulatory bodies, are essential to move ML-based disease prediction systems toward demonstrable real-world clinical utility, particularly in resource-constrained regions where specialist access remains limited.

VI. CONCLUSION

This review has synthesized fifteen representative studies spanning diabetes, cardiovascular disease, and Parkinson's disease prediction using machine learning, organized around disease domain and supplemented by an emerging body of multi-disease integration research. Across these studies, no single algorithm consistently dominates; rather, performance depends heavily on dataset characteristics, feature engineering choices, and validation methodology, with ensemble and hybrid approaches frequently, though not universally, outperforming standalone classifiers. Addressing the identified gaps—particularly dataset diversity, evaluation standardization, interpretability, and genuine multi-disease integration—represents the most direct path toward translating this substantial body of academic research into reliable, clinically deployable tools that can support earlier diagnosis and improved health outcomes worldwide.

REFERENCES

[1] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Comput. Sci., vol. 132, pp. 1578–1585, 2018.

[2] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017.

[3] M. Maniruzzaman et al., "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," Comput. Methods Programs Biomed., vol. 152, pp. 23–34, 2017.

[4] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," Front. Genet., vol. 9, art. 515, 2018.

[5] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," IEEE Access, vol. 8, pp. 76516–76531, 2020.

[6] M. M. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood prediction of diabetes at early-stage using data mining techniques," in Computer Vision and Machine Intelligence in Medical Image Analysis, Singapore: Springer, 2020, pp. 113–125.

[7] Y. K. Qawqzeh, A. S. Bajahzar, M. Jemmali, M. M. Otoom, and A. Thaljaoui, "Classification of diabetes using photoplethysmogram (PPG) waveform analysis: Logistic regression modeling," BioMed Res. Int., vol. 2021, art. 6663622, 2021.

[8] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," Am. J. Cardiol., vol. 64, no. 5, pp. 304–310, 1989.

[9] R. Alizadehsani et al., "A data mining approach for diagnosis of coronary artery disease," Comput. Methods Programs Biomed., vol. 111, no. 1, pp. 52–61, 2013.

[10] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019.

[11] D. Shah, S. Patel, and S. K. Bharti, "heart disease prediction using machine learning techniques," SN Comput. Sci., vol. 1, no. 6, art. 345, 2020.

[12] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," Comput. Intell. Neurosci., vol. 2021, art. 8387680, 2021.

[13] M. Swathy and K. Saruladha, "A

comparative study of classification and prediction of cardio-vascular diseases (CVD) using machine learning and deep learning techniques," ICT Express, vol. 8, no. 1, pp. 109–116, 2021.

[14] M. Peng, F. Hou, Z. Cheng, T. Shen, K. Liu, C. Zhao, and W. Zheng, "Prediction of coronary artery disease via XGBoost algorithm with Bayesian hyperparameter optimization," PLOS ONE, vol. 17, no. 10, art. e0275221, 2022.

[15] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, and I.
M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," BioMed. Eng. OnLine, vol. 6, no. 1, art. 23, 2007.

[16] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," IEEE Trans. Biomed. Eng., vol. 56, no. 4, pp. 1015–1022, 2009.

[17] M. Shahbakhi, D. T. Far, and E. Tahami, "Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine," J. Biomed. Sci. Eng., vol. 7, no. 4, pp. 147–156, 2014.

[18] S. Grover, S. Bhartia, A. Yadav, and K. R. Seeja, "Predicting severity of Parkinson's disease using deep learning," Procedia Comput. Sci., vol. 132, pp. 1788–1794, 2018.

[19] U. Ozkaya, L. Seyfi, and S. Ozturk, "Comparative analysis of machine learning algorithms for Parkinson's disease detection using voice measurements," J. Med. Syst., vol. 44, no. 7, art. 134, 2020.

[20] J. M. Tracy, Y. Özkanca, D. C. Atkins, and R. Hosseini Ghomi, "Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease," J. Biomed. Inform., vol. 104, art. 103362, 2020.

[21] G. T. Reddy, N. Khare, and S. Bhattacharya, "An integrated multi-disease prediction framework using machine learning for mobile health applications," Health Inform. J., vol. 29, no. 2, pp. 1–18, 2023.

[22] D. C. Yadav and S. Pal, "Prediction of multiple disease using ensemble machine learning approach for healthcare data," Int. J. Inf. Technol., vol. 14, no. 5, pp. 2517–2528, 2022.

[23] World Health Organization, "Noncommunicable diseases progress monitor 2022," WHO, Geneva, Switzerland, 2022.

[24] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: IDF, 2021.