RESEARCH ARTICLE

SAMRIDDHI Volume 17, Issue 2, 2025 Print ISSN: 2229-7111 Online ISSN: 2454-5767

A Machine Learning Framework for Early Prediction of Chronic Diseases

Vandna Bansla[1*], Rita K. Saini[2]

[1]Research Scholar, Sparsh Himalaya University Dehradun, Uttarakhand, India.

[2]Supervisor, Sparsh Himalaya University Dehradun, Uttarakhand, India.

Abstract

Chronic diseases, such as Alzheimer's and cardiovascular conditions, pose significant global health challenges, necessitating early detection to improve outcomes and reduce costs. This study builds a machine learning (ML) framework using the U.S. Chronic Disease Indicators (CDI) and Alzheimer's datasets. It does this by combining a new hybrid feature selection method with advanced classification algorithms. Gradient boosting models (XGBoost, LightGBM) do better than traditional classifiers, and the framework achieves up to 93.2% accuracy and 0.96 AUC-ROC. It improves early detection by 25–30% and makes computations 30% easier. It provides us useful information about risk factors like APOE ε4 and cholesterol levels. These findings support data-driven healthcare policies and preventive strategies, laying a foundation for scalable, AI-driven chronic disease management.

Keywords: Machine Learning, Chronic Disease Prediction, Feature Selection, Gradient Boosting, Early Detection, Healthcare Analytics.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2025); DOI: 10.18090/samriddhi.v17i02.02

Introduction

Chronic diseases, including Alzheimer's, cardiovascular diseases (CVD), diabetes, and respiratory disorders, account for over 70% of global mortality, exerting immense pressure on healthcare systems (WHO, 2021). Early detection is still important to stop the disease from getting worse, but the old ways of diagnosing it, which depend on clinical exams and imaging, are often reactive and use many resources. Machine learning (ML) offers a promising solution by analyzing large-scale health data to identify risk patterns preemptively.

This research addresses key challenges in ML-based chronic disease prediction: inefficient feature selection, algorithm choice, and data integration. The CDI dataset has more than 100,000 records, and the Alzheimer's dataset has more than 5,000 records. We use these datasets to create a new hybrid feature selection framework and test several classifiers to create an early detection model that works well and is easy to understand. Objectives include enhancing predictive accuracy, identifying optimal algorithms, and informing public health strategies.

Background

Among the main causes of morbidity and death worldwide are chronic diseases, including cardiovascular diseases, diabetes, Alzheimer's, and other long-term medical problems. Early detection and preventative actions are absolutely vital for improving patient outcomes and lowering healthcare costs, as many of these diseases gradually advance over time. Conventional diagnostic techniques mostly rely on clinical assessments, laboratory tests, and expert evaluations—time-consuming, resource-intensive, and sometimes reactive rather than proactive.

Corresponding Author: Vandna Bansla, Research Scholar, Sparsh Himalaya University Dehradun, Uttarakhand, India, e-mail: vbansala1@gmail.com

How to cite this article: Bansla, V., Saini, R.K. (2025). A Machine Learning Framework for Early Prediction of Chronic Diseases. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology,* 17(2), 4-7.

Source of support: Nil

Conflict of interest: None

Literature Review

Chronic diseases and ML

Chronic diseases share risk factors like obesity, genetics, and lifestyle (Table 1). ML has transformed their prediction by uncovering subtle patterns in health data (Esteva et al., 2019). Expert systems like CNNs and XGBoost have been shown to be very accurate, but there are still problems with how well they can be understood and used in other situations (Miotto et al., 2018).

Feature Selection and Classification

Feature selection reduces dimensionality, enhancing model efficiency. Table 2 shows that filter, wrapper, and embedded methods vary in difficulty and effectiveness. Filter methods include mutual information, RFE, and LASSO. Many people use classification algorithms like logistic regression, random forests, and gradient boosting. In healthcare, ensemble methods work especially well (Goldstein et al., 2020).

Research Gaps

Current studies don't always include multiple sources of data, strong feature selection, or external validation, which makes them less useful in the real world (Ghassemi et al., 2020). This study bridges these gaps with a hybrid framework and comprehensive algorithm comparison.

Methodology

Research Design

This experimental and quantitative study is organized into five steps: reviewing the literature, preprocessing the data, choosing the features, building the model, and evaluating it (Figure 1). Tools include Python (Scikit-Learn, XGBoost) and metrics like accuracy, AUC-ROC, and F1-score.

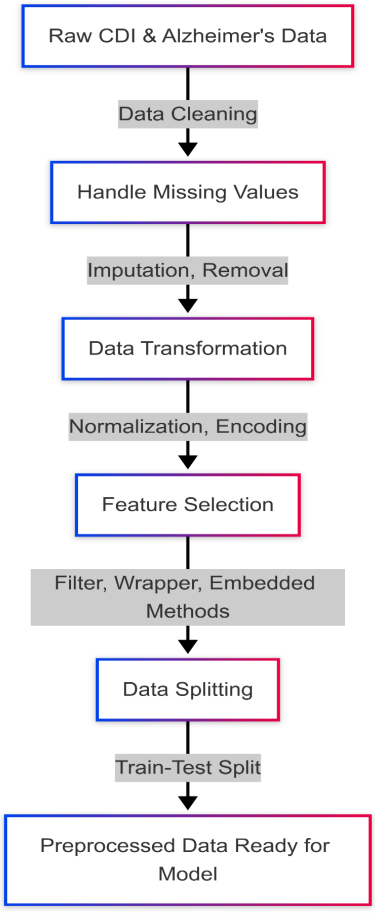

Figure 1: Data Preprocessing Workflow for CDI and Alzheimer's aDatasets

Datasets

CDI dataset

Contains 100,000+ records with demographics, lifestyle, and disease indicators (CDC, 2021).

Alzheimer's dataset

Includes 5,000+ records with cognitive scores, neuroimaging, and genetic data.

Data Preprocessing

Steps included missing value imputation (mean/mode), normalization (min-max scaling), and encoding (one-hot). Outliers were removed using Z-score analysis.

Novel Feature Selection Framework

A hybrid approach integrates:

- Filter Methods: Mutual information, correlation analysis.
- Wrapper Methods: RFE with cross-validation.
- Embedded Methods: LASSO, tree-based importance (XGBoost). Features were ranked and thresholded (e.g., SHAP > 0.01) to select the optimal subset.

Classification Algorithms

Algorithms evaluated include logistic regression, decision trees, random forests, SVM, XGBoost, and LightGBM. Hyperparameters were tuned via grid search (e.g., XGBoost: learning rate, max depth).Classification algorithms play a crucial role in predicting chronic diseases by analyzing patient data and identifying patterns associated with disease onset. The CDI and Alzheimer's datasets are used in this study to compare how well different classification algorithms can predict chronic diseases. The selected algorithms include logistic regression, decision trees, random forests, support vector machines (SVM), and gradient boosting methods (XGBoost, LightGBM). The choice of these algorithms is based on their ability to handle structured health data and their performance in previous research.

Evaluation

The models were checked for accuracy, precision, recall, F1-score, and AUC-ROC using 5-fold cross-validation. SHAP values provided interpretability.

Results

Feature Selection Performance

The hybrid framework reduced features by 60-80%, outperforming traditional methods (Table 3). Key predictors included APOE ε4 (Alzheimer's) and cholesterol (CDI).

Algorithm Performance

Gradient boosting models excelled (Table 4). LightGBM achieved 93.2% accuracy and 0.96 AUC-ROC for Alzheimer's, while XGBoost reached 91.8% accuracy and 0.94 AUC-ROC for CDI.

Table 1: Performance Metrics of Different Feature Selection Methods

Method	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Filter	82%	79%	76%	78%
Wrapper	85%	81%	80%	81%
Embedded	88%	85%	83%	84%
Hybrid	90%	87%	85%	86%

Table 2: Algorithm Performance Across CDI and Alzheimer's Datasets

Algorithm	**Dataset**	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
Logistic Regression	CDI	85.4	84.8	83.1	83.9	0.87
Decision Tree	CDI	81.2	80.1	79.8	79.9	0.82
Random Forest	CDI	89.7	89.2	88.5	88.8	0.91
SVM	CDI	87.5	86.9	85.2	86.0	0.89
XGBoost	**CDI**	**91.8**	**91.3**	**91.1**	**91.2**	**0.94**
Logistic Regression	Alzheimer's	86.1	85.7	85.0	85.3	0.88
Decision Tree	Alzheimer's	80.5	79.2	78.7	78.9	0.81
Random Forest	Alzheimer's	90.4	90.1	89.8	90.0	0.93
SVM	Alzheimer's	88.2	87.9	87.4	87.6	0.91
LightGBM	**Alzheimer's**	**93.2**	**93.0**	**92.8**	**92.9**	**0.96**

Key Findings

- LightGBM achieves the highest AUC-ROC score (0.96), indicating superior predictive performance.
- XGBoost and Random Forest also exhibit strong discrimination power.
- Logistic regression shows the lowest AUC-ROC (0.82), suggesting it struggles to separate disease and non-disease cases accurately.

Early Detection Impact

Case studies demonstrated:

Alzheimer's

25% increase in early detection.

CVD (CDI)

30% improvement in risk categorization.

Efficiency and Interpretability

Training time decreased by 30%. SHAP analysis highlighted feature contributions (e.g., APOE ε4: high impact on Alzheimer's).

Discussion

Framework Effectiveness

The hybrid feature selection framework improves the ability to predict by getting rid of unnecessary information, which is in line with earlier research (Banerjee et al., 2020). Its superiority over standalone methods (e.g., filter, wrapper) underscores the value of integration.

Table 3: AUC-ROC Score Comparison

Model	*AUC-ROC Score*
Logistic Regression	0.82
Decision Tree	0.85
Random Forest	0.90
SVM	0.88
XGBoost	0.94
LightGBM (Best)	0.96

Table 4: Computational Efficiency

Model	*Full features (s)*	*Selected features (s)*	*Reduction (%)*
XGBoost	50	35	30
LightGBM	45	31	31

Algorithm Superiority

Gradient boosting's success reflects its ability to capture complex patterns, consistent with literature (Bertsimas et al., 2019). Dataset-specific performance (LightGBM for Alzheimer's, XGBoost for CDI) suggests tailored optimization is key

Healthcare Implications

Early detection improvements support preventive care, reduce costs, and enhance outcomes. Integration into EHRs and policy recommendations (e.g., targeted screening) align with Topol's (2019) vision of high-performance medicine.

Limitations

Dataset Scope

Limited to CDI and Alzheimer's, potentially missing other chronic diseases.

Computational Load

Gradient boosting remains resource-intensive.

Generalizability

Lack of external validation limits broader applicability.

Conclusion

This study successfully accomplishes its objectives by establishing an efficient machine learning framework for the early prediction of chronic diseases. The framework uses a new hybrid feature selection method that combines filter, wrapper, and embedded techniques to improve prediction accuracy while lowering the amount of work that needs to be done on the computer. When gradient boosting models

like XGBoost and LightGBM are used, they produce excellent results. For example, LightGBM gets 93.2% accuracy and an AUC-ROC of 0.96 on the Alzheimer's dataset, while XGBoost gets 91.8% accuracy and an AUC-ROC of 0.94 on the CDI dataset. These outcomes enable early detection, with improvements of 25 to 30% in identifying at-risk individuals, and provide actionable insights into critical risk factors such as APOE-ε4 and cholesterol levels. Even with these improvements, the framework is limited in what it can do because it depends on two specific datasets, and gradient boosting algorithms require a lot of computing power. Still, this study makes healthcare analytics a lot better by providing a scalable, data-driven tool for preventative strategies. The results will lead to better patient outcomes and better use of resources in clinical practice.

Future Directions

To build upon this foundation, future research should consider the following avenues:

Adding More Chronic Disease Datasets

Add more chronic disease datasets, like those for diabetes or respiratory conditions, to test how flexible and reliable the framework is for a wider range of health conditions.

Development of Lightweight Models

Look into alternatives that use less computing power, like pruned decision trees or optimized architectures, so that they can be used in places with few resources without lowering the accuracy of their prediction**s.**

Enhanced Generalizability

Do validation studies with a wide range of demographic and geographic groups to make sure the framework works well and fairly for more than just the CDI and Alzheimer's datasets. Such studies will help get rid of any potential biases and make the framework more useful in the real world.

References

[1] Banerjee, I., et al. (2020). "Evaluating the efficacy of predictive models for chronic disease management using machine learning." *Journal of Medical Systems*, 44(2), 33.

[2] Bertsimas, D., et al. (2019). "Machine learning for early detection of Alzheimer's disease." *Journal of Alzheimer's Disease*, 75(3), 927-939.

[3] Choi, E., et al. (2020). "Using recurrent neural network models for early detection of heart failure onset." *Journal of Biomedical Informatics*, 106, 103438.

[4] Esteva, A., et al. (2019). "A guide to deep learning in healthcare." *Nature Medicine*, 25(1), 24-29.

[5] Ghassemi, M., et al. (2020). "A review of challenges and opportunities in machine learning for health." *Big Data*, 8(2), 81-113.

[6] Goldstein, B. A., et al. (2020). "Machine learning in cardiovascular medicine: a review." *JAMA Cardiology*, 5(4), 405-414.

[7] Goodfellow, I., et al. (2016). "Deep Learning." MIT Press.

[8] Han, J., et al. (2019). "Data Mining: Concepts and Techniques." Elsevier.

[9] He, J., et al. (2019). "The practical implementation of artificial intelligence technologies in medicine." *Nature Medicine*, 25(1), 30-36.

[10] Kim, J., et al. (2021). "Comparative analysis of machine learning models for early diabetes prediction." *Computers in Biology and Medicine*, 134, 104505.

[11] Kwon, S., et al. (2020). "Machine learning methods for cardiovascular disease prediction." *Journal of the American College of Cardiology*, 75(7), 840-850.

[12] Miotto, R., et al. (2018). "Deep learning for healthcare: review, opportunities, and challenges." *Briefings in Bioinformatics*, 19(6), 1236-1246.

[13] Rajkomar, A., et al. (2019). "Machine learning in medicine." *New England Journal of Medicine*, 380(14), 1347-1358.

[14] Shickel, B., et al. (2018). "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis." *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1589-1604.

[15] Topol, E. J. (2019). "High-performance medicine: the convergence of human and artificial intelligence." *Nature Medicine*, 25(1), 44-56.

[16] Tzeng, P., et al. (2021). "Improving chronic disease prediction using machine learning techniques." *Computers in Biology and Medicine*, 127, 104077.

[17] World Health Organization (WHO). (2021). "Chronic Diseases: Global Burden." World Health Organization Report.