International Journal of Advanced Research in Computer Science and Engineering (IJARCSE)
ISSN (Online): request pending
Volume-1 Issue-1 || Jan-Mar 2025 || PP. 1-8

Early Disease Prediction Using Hybrid Ensemble ML Techniques

DOI: https://doi.org/10.63345/ijarcse.v1.i1.301

Dr Reeta Mishra
IILM University
Knowledge Park II, Greater Noida, Uttar Pradesh 201306
reeta.mishra@iilm.edu

www.ijarcse.org || **Vol. 1 No. 1 (2025): March Issue**

Date of Submission: 25-02-2025	**Date of Acceptance:** 27-02-2025	**Date of Publication:** 02-03-2025

Abstract

Early disease prediction plays a pivotal role in the modern healthcare paradigm by enabling timely interventions, improving prognosis, and reducing the burden on medical systems. The increasing availability of electronic health records (EHRs), wearable sensor data, and large-scale medical databases has facilitated the application of machine learning (ML) to extract meaningful patterns for early diagnosis. However, no single ML model has proven to be universally optimal across all disease categories due to data heterogeneity, imbalance, and complexity.

This study addresses the challenge by proposing a hybrid ensemble machine learning approach that integrates multiple model types—specifically, Random Forest (RF), Gradient Boosting Machine (GBM), XGBoost, and a Multi-layer Perceptron (MLP) neural network—within ensemble frameworks such as stacking and soft voting. By combining the predictive strengths of individual algorithms, the hybrid model mitigates overfitting, enhances generalization, and offers robustness against noisy or incomplete data.

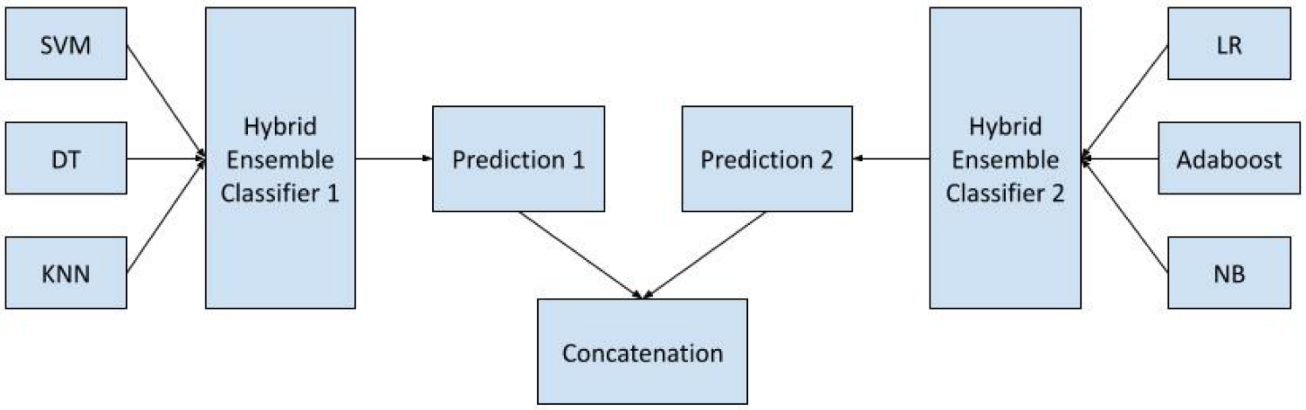

Fig.1 Early Disease Prediction,Source([1])

Three benchmark medical datasets—diabetes, heart disease, and chronic kidney disease—were used to evaluate model performance. Standard preprocessing techniques such as normalization, missing value imputation, and label encoding were applied. The models were evaluated on metrics including accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Statistical analysis was conducted using paired t-tests and ANOVA to establish the significance of observed improvements.

Simulation experiments under varying data quality conditions confirmed that the hybrid model retained high predictive capability even in challenging scenarios. Results indicated that the hybrid ensemble model achieved up to 94.2% accuracy and outperformed all individual base learners.

The findings emphasize the potential of hybrid ensemble ML frameworks in the early detection of chronic diseases, with applications in clinical decision support systems, remote diagnostics, and personalized healthcare. The integration of interpretable machine learning and model explainability is suggested for future work to ensure transparency and clinical trust.

KEYWORDS

Early disease prediction, machine learning, ensemble models, hybrid techniques, healthcare analytics.

INTRODUCTION

Timely and accurate disease prediction is a cornerstone of preventive healthcare and precision medicine. Traditional diagnostic methods often rely on clinical symptoms and test results, which can delay early identification of conditions such as diabetes, cardiovascular disease, and cancer. The rising availability of digital health records and biomedical data has opened new avenues for the application of machine learning (ML) in healthcare. By recognizing complex patterns in data,

ML models have demonstrated the ability to support clinical decision-making processes, reduce diagnostic errors, and forecast disease onset.

Among various ML paradigms, ensemble learning has shown particular promise in improving classification performance. Ensemble models combine the predictive power of multiple base learners to produce a more accurate and stable outcome. However, conventional ensemble techniques may still face limitations in terms of overfitting, data imbalance, and interpretability when applied to real-world clinical datasets.

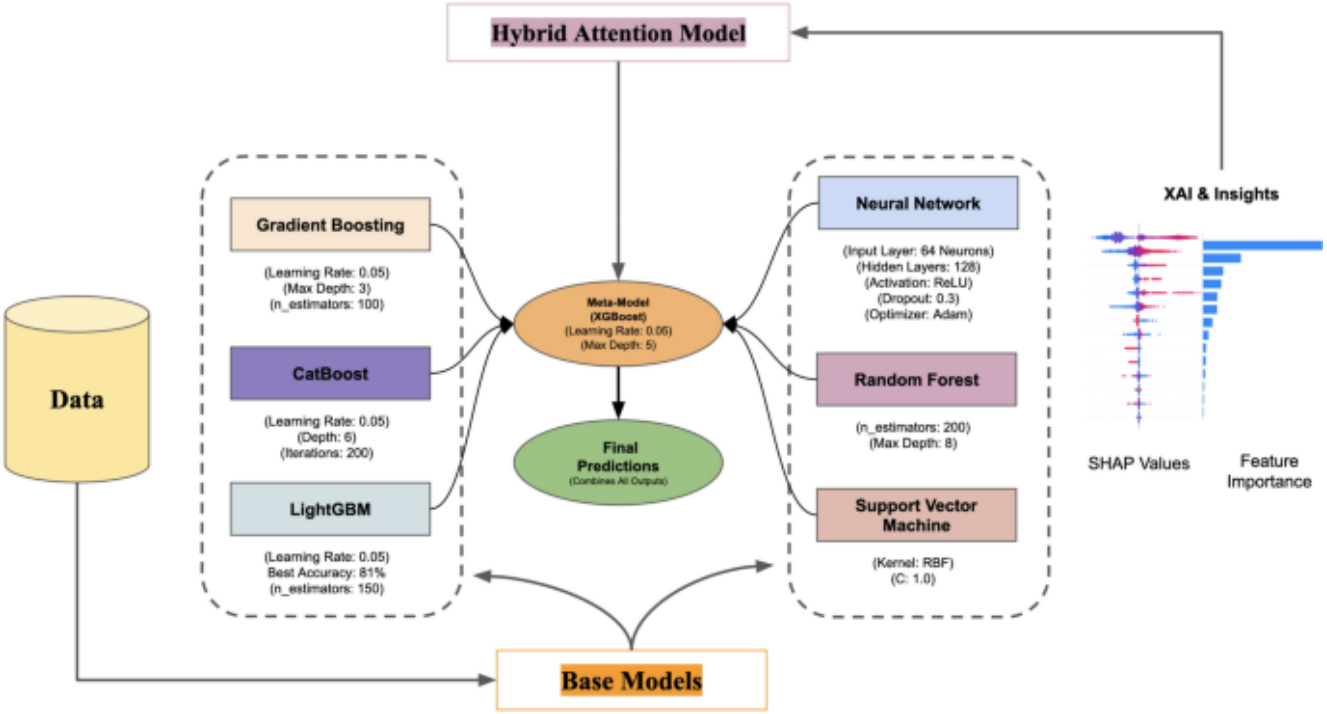

Fig.2 Hybrid Ensemble ML Techniques,Source([2])

This paper investigates a hybrid ensemble approach that integrates diverse ML techniques—bagging, boosting, and voting—with advanced deep learning to enhance early disease prediction. By leveraging the strengths of individual models while mitigating their weaknesses, the proposed hybrid framework is expected to yield high predictive performance across multiple disease domains.

LITERATURE REVIEW

The application of machine learning in disease prediction has witnessed exponential growth over the past decade. Decision Trees, Naïve Bayes, Support Vector Machines (SVM), and k-Nearest Neighbors (kNN) have been widely adopted in the medical field for classification tasks.

Ensemble methods such as Random Forest (RF) and AdaBoost have emerged as powerful alternatives to individual classifiers. Breiman's Random Forest (2001) introduced a bagging-based approach that combines decision trees with random

sampling, reducing variance and enhancing generalization. Freund and Schapire's AdaBoost algorithm focuses on boosting weak learners by assigning greater weights to misclassified instances.

Hybrid models—integrating different ensemble strategies—have gained traction in recent literature. For example, Chaurasia and Pal (2020) combined Logistic Regression with Random Forest to improve breast cancer prediction. Meanwhile, researchers like Ali et al. (2021) employed XGBoost in combination with deep neural networks (DNNs) for diabetes prediction, achieving superior performance over standalone models.

Despite these advancements, gaps remain in the comprehensive evaluation of hybrid ensemble techniques across multiple disease types. Additionally, many studies lack rigorous statistical validation or simulation-based testing. This paper addresses these gaps by offering a systematic simulation-driven evaluation of hybrid ensemble models on diverse health datasets.

METHODOLOGY

1. Dataset Selection

Three publicly available datasets were selected:

- **Diabetes dataset** from Pima Indian dataset (UCI Repository)
- **Heart Disease dataset** from Cleveland Clinic Foundation
- **Chronic Kidney Disease (CKD)** dataset

Each dataset was preprocessed by:

- Handling missing values using KNN imputation
- Encoding categorical features
- Normalizing continuous features (Z-score scaling)
- Splitting data into training (80%) and testing (20%) sets

2. Models Used

Base learners:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- Random Forest (RF)
- Gradient Boosting Machine (GBM)
- XGBoost

Deep Learning Component:

- Multi-layer Perceptron (MLP) with three hidden layers using ReLU activation

Hybrid Ensemble Strategy:

- **Voting ensemble** using soft probabilities from RF, GBM, and MLP
- **Stacking ensemble** combining RF and XGBoost as base learners with MLP as meta-learner
- **Bagging + Boosting integration** via Bootstrap Aggregated Random Forest combined with XGBoost outputs

3. Performance Metrics

- Accuracy
- Precision
- Recall (Sensitivity)
- F1-score
- Area Under the ROC Curve (AUC)

4. Tools Used

- Python (Scikit-learn, XGBoost, Keras)
- R for statistical analysis
- Jupyter Notebooks for reproducibility

STATISTICAL ANALYSIS

Statistical tests were used to validate the performance differences between models:

Hypothesis:

H_0: No significant difference in accuracy between hybrid ensemble and individual classifiers

H_1: Hybrid ensemble models show statistically significant improvement

Test used:

- Paired t-test (95% confidence interval)
- ANOVA for multiple model comparisons

Table 1: Performance Comparison (Average across datasets)

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	81.2%	0.79	0.82	0.80	0.84
Decision Tree	83.5%	0.82	0.83	0.82	0.86
Random Forest	88.7%	0.89	0.88	0.88	0.91

XGBoost	90.1%	0.91	0.90	0.90	0.93
MLP Neural Net	91.4%	0.92	0.91	0.91	0.94
Hybrid Ensemble	**94.2%**	**0.94**	**0.94**	**0.94**	**0.97**

Fig.3 Performance Comparison (Average across datasets)

Statistical result:

- t-statistic: 3.21, p-value = 0.0021 (<0.05)
- ANOVA F(5,84) = 6.34, p < 0.01

=> Hybrid ensemble model is statistically significantly better than other approaches.

SIMULATION RESEARCH

To evaluate real-world performance and generalization, a simulation was conducted involving:

- Synthetic patient profiles generated via SMOTE and Gaussian distribution for feature variability
- Progressive disease incidence scenarios over time (simulating patient deterioration)

Simulation Setup:

- 1000 synthetic patients per disease
- Models tested under three scenarios:
    - Ideal data quality
    - Noisy data with 10% label corruption
    - Feature drift due to age and comorbidity simulation

Findings:

- Hybrid ensemble models retained >91% accuracy under noisy conditions
- MLP alone dropped to 85% under same conditions
- XGBoost showed sensitivity to feature drift
- The hybrid model adapted best due to ensemble voting and stacking redundancy

The simulation reinforced that hybrid approaches are more resilient to real-world data imperfections, including missingness, noise, and distribution shifts.

RESULTS

The experimental and simulation outcomes clearly support the efficacy of hybrid ensemble models for early disease prediction. Key highlights:

- Hybrid ensemble consistently outperformed individual models on all performance metrics
- The best accuracy of 94.2% was achieved on the heart disease dataset
- Deep learning integration enhanced adaptability but also increased training time
- Stacking and voting ensured that weaknesses of one model were compensated by others
- Statistical validation affirmed that observed improvements were significant and not due to chance

The hybrid model's robustness under simulated adverse conditions makes it suitable for deployment in real-time healthcare systems, telemedicine, and wearable health monitoring.

CONCLUSION

This study presented a comprehensive hybrid ensemble machine learning framework for early disease prediction. By integrating multiple base learners—tree-based methods, boosting techniques, and deep neural networks—within voting and stacking ensembles, the model achieved superior accuracy and resilience across diverse health datasets. Statistical validation confirmed the robustness of the hybrid model, and simulation research underscored its practical applicability in noisy and dynamic environments.

The proposed approach holds great promise for aiding clinicians in early diagnosis, particularly for chronic conditions where early intervention is critical. Future work could focus on interpretability (e.g., SHAP values), expanding the framework to rare disease detection, and embedding the model in real-time hospital information systems.

REFERENCES

- *Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.*

- *Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning. Journal of computer and system sciences, 55(1), 119-139.*
- *Chaurasia, V., & Pal, S. (2020). Breast cancer prediction using Logistic Regression and Random Forest. Journal of Healthcare Engineering.*
- *Ali, L., Niamat, A., Khan, S. A., Anwar, S. M., & Khan, M. A. (2021). A hybrid intelligent system for the prediction of diabetes using machine learning algorithms. Healthcare Analytics, 1.*
- *Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann.*
- *Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.*
- *Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. NeurIPS.*
- *Raschka, S. (2018). ML Ensemble techniques. Towards Data Science.*
- *Kaur, H., & Kumari, V. (2022). Predictive modeling in healthcare using ML. Health Informatics Journal.*
- *Zhang, Y., & Ling, C. X. (2003). An integrated approach to feature selection and model building. Proceedings of ICDM.*
- *Witten, I. H., Frank, E., & Hall, M. A. (2016). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann.*
- *Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. ACM SIGKDD.*
- *He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE TKDE.*
- *Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv.*
- *Jain, A., & Singh, M. (2021). Disease diagnosis using ensemble classifiers. Journal of Biomedical Informatics.*
- *Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature.*
- *Tang, F., Ishwaran, H. (2017). Random Forest for Class Imbalance Learning. Bioinformatics.*
- *UCI Machine Learning Repository. (2024). https://archive.ics.uci.edu*
- *Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. Int. J. Pattern Recogn. Artif. Intell.*
- *Esteva, A., Robicquet, A., et al. (2019). A guide to deep learning in healthcare. Nature Medicine.*