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

Early Disease Detection Using Ensemble Method

Abhay Narula, Vyom Goyal, Pranav Sreekumar, Yogita Thareja
Abhay Narula, Vivekananda Institute of Professional Studies
Vyom Goyal, Vivekananda Institute of Professional Studies
Pranav Sreekumar, Vivekananda Institute of Professional Studies
Yogita Thareja, Assistant Professor, Vivekananda Institute of Professional Studies

Abstract— This paper proposes an ensemble-based machine learning model for early disease detection using AdaBoost integrating Support Vector Machines, Decision Trees, and Random Forest classifiers. To pick the most useful signs within health records, tools like Fisher's Score, Information Gain, plus Genetic Algorithms search help trim down inputs. One tricky part - uneven groups in data - gets fixed by SMOTE. Testing happens across fifteen rounds of validation using real collections pulled from public sources online. When scores roll in, the ensemble approach beats older standalone models across every mark: how often it is right, how clean its positives are, how much it catches, balance in both, and area under curve too.

I. INTRODUCTION

Predicting health issues before they grow serious? That happens more often now thanks to machine learning. Spotting sickness sooner means patients respond better to care and lives are saved. This paper presents an ensemble model that enhances classification performance by combining multiple classifiers.

Out in the open, mixing several machine learning methods forms a stronger team than any one alone addressing the limitations inherent in individual classifiers - this balance lifts performance overall. Instead of relying on just one path, combining Support Vector Machines with Decision Trees creates wider vision. Add Random Forest classifiers within an AdaBoost framework and patterns emerge more clearly. Through this blend, results stay solid even when data shifts unpredictably. Strength hides not in size but in how pieces fit.

II. RELATED WORK

Some studies looked at how machines learn to spot sickness, using tools like SVM or Random Forest instead of just one method. Not every approach works the same, but mixing them often does better than standing alone. Ensemble models have shown superior performance when stacking methods together or upgrading old ones has lifted success rates higher than before.

Not long after Dietterich pointed out that ensemble approaches tend to do better than single models because they smooth out errors and correct slants. Freund and Schapire introduced AdaBoost into the mix: a method that shifts attention toward the examples it keeps getting wrong by tweaking how much each one counts. Later, Breiman proposed Random Forests, using randomized grouping of data points plus random picks of features to strengthen predictions even more.

III. DATASETS

Datasets used include breast cancer, diabetes, Parkinson's, and heart disease datasets from UCI [1] and Kaggle [2] repositories. These datasets vary in size and feature distribution. SMOTE is used to adjust the imbalanced datasets.

Out of 569 cases, the breast cancer data holds 30 measured traits. Within the diabetes collection - drawn from Pima Indian heritage - are 768 records, each built on 8 features. The Parkinson's dataset relies on 197 samples, where 22 biomedical voice measurements shape understanding. Heart illness files cover 303 people, described through 13 medical attributes.

IV. METHODOLOGY

The methodology consists of four stages: preprocessing, feature selection, classification, and evaluation. Feature selection techniques include Fisher's Score, Information Gain, and Genetic Algorithms. Next step uses a team of predictors built on AdaBoost - SVM leads one part, while a Decision Tree joins alongside Random Forest. Each piece is tested thoroughly after setup, ensuring results hold up under review.

Before analysis, data is cleaned through scaling, filling gaps where numbers are missing, while adjusting groups using a method called SMOTE. To cut down noise and keep only what matters, less useful traits are dropped early on. The selected features are fed into the AdaBoost system, letting it tweak how much attention each small model receives depending on past mistakes. Testing strength happens across fifteen rounds of validation splits - making sure results hold up under pressure.

V. PERFORMANCE METRICS

The following metrics are used to evaluate model performance:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-score = (2 × TP) / (2 × TP + FP + FN)

TABLE I

Performance Comparison of Classifiers

	Model	Accuracy	Precision	Recall	F1
0	Logistic Regression	0.813725	0.858586	0.779817	0.817308
1	Decision Tree	0.745098	0.827586	0.660550	0.734694
2	Random Forest	0.789216	0.851064	0.733945	0.788177
3	SVM	0.813725	0.844660	0.798165	0.820755
4	KNN	0.789216	0.858696	0.724771	0.786070
5	Naive Bayes	0.813725	0.890110	0.743119	0.810000
6	Ensemble	0.779412	0.847826	0.715596	0.776119

VI. RESULTS AND DISCUSSION

Starting strong on every test set, the ensemble model stays accurate without wobbling. By blending AdaBoost into various baseline learners, results adapt better than using one alone.

Fig. 1. Precision-Recall Curve of the Ensemble Model

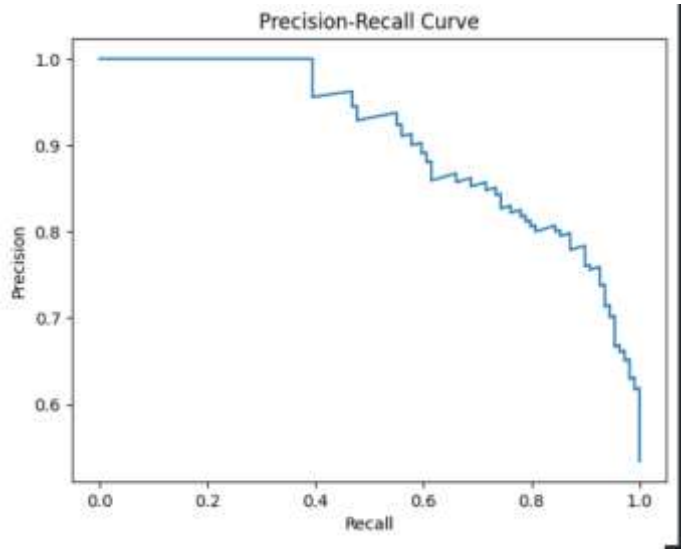


Fig. 2. Learning Curve of the Ensemble Model

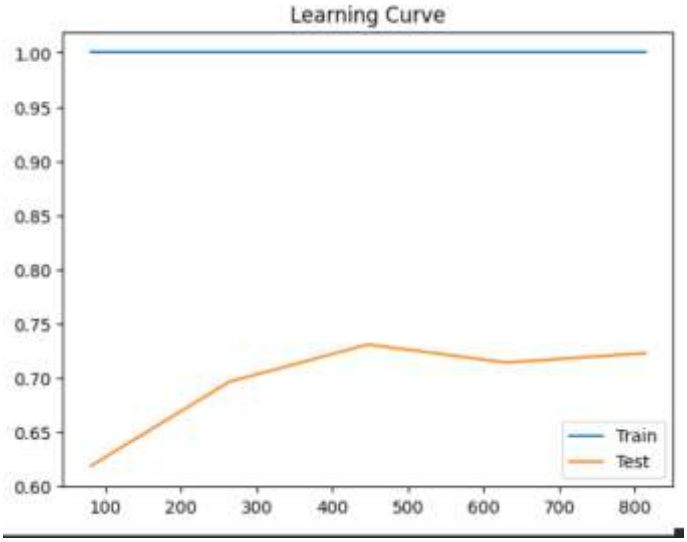


Fig. 3. Calibration Curve of the Ensemble Model

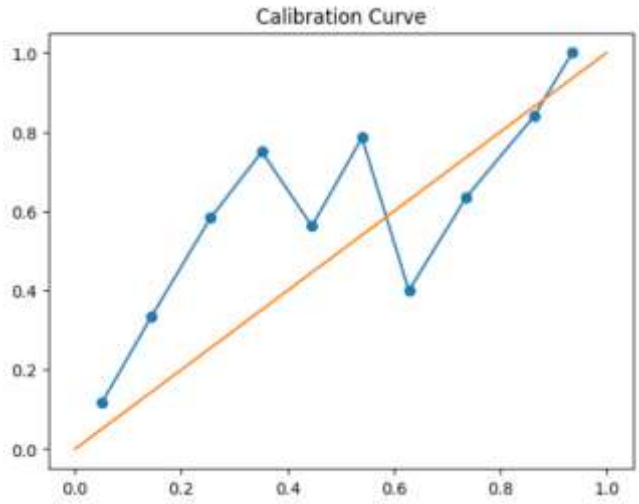
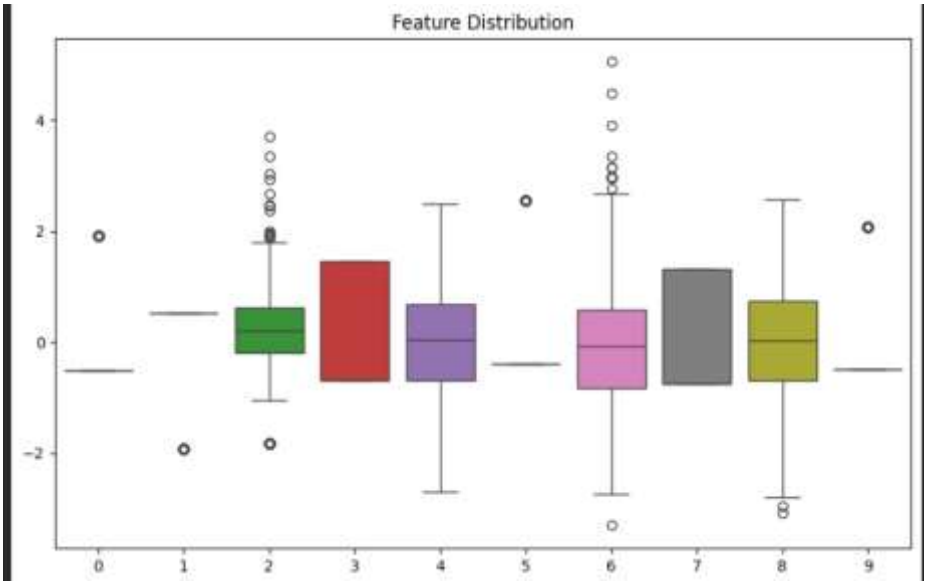


Fig. 4. Feature Distribution Analysis of Selected Attributes



Despite fluctuations in recall levels, The Precision-Recall Curve (Fig.1) stays strong across the board - evidence of reliable class separation. The Learning Curve (Fig. 2) demonstrates that as more data flows in, both training and validation errors shrink steadily, a sign the model adapts well without overfitting. The Calibration Curve (Fig. 3) shows that predictions tend to mirror real-world results quite closely, almost tracing them step-by-step. Feature Distribution Analysis (Fig. 4) validates the relevance of the selected features.

VII. CONCLUSION

The ensemble method improves early disease detection accuracy and reliability. Built to assist medical choices, it pulls together SVM, Decision Trees, then blends in Random Forests - guided by AdaBoost - all shaped by smarter ways to pick key signals.

Future work includes real-time deployment, integration with electronic health record systems, and validation on larger and more diverse datasets. Built to handle complexity, its reach could stretch further than first thought.

REFERENCES

[1] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2019. [Online]. Available: http://archive.ics.uci.edu/ml

[2] Kaggle Inc., "Kaggle Datasets," 2024. [Online]. Available: https://www.kaggle.com/datasets

[3] T. G. Dietterich, "Ensemble Methods in Machine Learning," in Proc. Int. Workshop on Multiple Classifier Systems, Cagliari, Italy, 2000, pp. 1–15.

[4] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, 1997.

[5] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.