Enhanced Prediction of Non-Communicable Diseases Using Ensemble Machine Learning Approaches

Non-communicable diseases (NCDs) are a major health issue worldwide and early detection is the key towards prevention and management. Conventional predictive models cannot always reflect complex, non-linear association between anthropometric indicators, and this limitation restricts their clinical application. This paper explores the effectiveness of ensemble machine learning (ML) methods, such as Random Forest, AdaBoost, CatBoost and XGBoost, in predicting NCDs on anthropometrics of adult subjects in the Jaffna teaching hospital and Sabaragamuwa University. The characteristics vital in this part were age, Body mass index, waist-hip ratio, visceral fat area, body fat percentage, total body water and muscle body fat. Compared to the traditional classifiers like Support Vector Machine, Artificial Neural Network, Logistic Regression, and Decision Tree, the performance comparisons with the ensemble models showed that they always surpassed the traditional ones, reaching up to 98.9 percent accuracy. The most significant predictors were visceral fat area, BMI and the waist-hip ratio. The results highlight the high potential of ensemble ML methods to deliver powerful, precise and scalable instruments in assessing early risks of NCDs, which justifies their integration into clinical decision-support systems.

Keywords: *Non-communicable diseases, ensemble machine learning, Random Forest, XGBoost, anthropometric indicators, predictive modeling*

Introduction

Non-communicable diseases (NCDs) represent a growing global health burden, contributing significantly to mortality and reduced quality of life. Traditional predictive models often struggle to capture the complex, non-linear patterns associated with NCD risk, limiting their effectiveness in clinical practice. Machine learning (ML) approaches have gained attention for their ability to model such complexities, and ensemble methods in particular have shown strong promise. By combining multiple weak or base learners into a single, more robust model, ensemble techniques reduce overfitting and enhance predictive performance. This study aims to evaluate the effectiveness of ensemble machine learning approaches in predicting NCDs using anthropometric data and to compare their performance against individual traditional ML models.

Materials and Methods

The dataset for this study was derived from adult participants aged over 18 years at Jaffna Teaching Hospital (JTH) and Sabaragamuwa University (SUSL), excluding children and pregnant women to avoid bias in body composition metrics. The features collected included age, weight, height, gender, muscle body fat (MBF), total body water (TBW), percentage body fat (PBF), body mass index (BMI), visceral fat area (VFA), and waist–hip ratio (WHR). Data preprocessing steps involved cleaning, normalization, and handling missing values to ensure data quality, as well as encoding categorical variables such as gender for compatibility with ML algorithms. Four ensemble models Random Forest (RF), AdaBoost, CatBoost, and Extreme Gradient Boosting (XGBoost) were implemented and compared with conventional approaches, including Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistic Regression, and Decision Tree. The target variable was defined as the presence or absence of NCDs.

Results

The performance evaluation demonstrated that ensemble models consistently outperformed traditional classifiers in predicting NCDs. Random Forest achieved the highest accuracy of 98.9%, closely followed by XGBoost and ANN with accuracies of 97.8%. Boosting techniques such as AdaBoost and CatBoost also showed strong performance, confirming the advantage of combining multiple learners to build reliable predictors. In contrast, traditional models such as Logistic Regression and SVM produced comparatively lower accuracies of 88.5% and 85.2%, respectively, highlighting their limitations in handling complex anthropometric interactions. Ensemble approaches not only improved accuracy but also provided robustness by reducing overfitting and enhancing generalizability. Among the anthropometric features, VFA, BMI, and WHR were consistently ranked as the most influential predictors of NCD presence. These findings suggest that ensemble learning can provide reliable and scalable solutions for early risk assessment and screening of NCDs in clinical practice.

Conclusion

This study confirms that ensemble machine learning approaches significantly improve the prediction of non-communicable diseases compared to traditional individual classifiers. Random Forest, along with boosting algorithms such as XGBoost and AdaBoost, demonstrated high accuracy and reliability when applied to anthropometric data, underscoring their potential for healthcare applications. The results highlight the importance of VFA, BMI, and WHR as critical features for NCD prediction, further supporting their clinical relevance. In the future, integrating these models into decision-support systems and testing them on larger, multi-center datasets could enhance their applicability and contribute to scalable, low-cost solutions for preventive healthcare.

References

[1] E. Maini, B. Venkateswarlu, B. Maini, and D. Marwaha, “Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India” Medical Journal Armed Forces India, vol. 77, pp. 302-311, 2021.

[2] S. M. S. Islam, A. Talukder, M. A. Awal, M. M. U. Siddiqui, M. M. Ahamad, B. Ahammed, et al., “Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries” Frontiers in Cardiovascular Medicine, vol. 9, 2022.

[3] M. A. Nematollahi, A. Askarinejad, A. Asadollahi, M. Salimi, M. Moghadami, S. Sasannia, et al., “Association and predictive capability of body composition and diabetes mellitus using artificial intelligence: a cohort study” 2022.

[4] K. Tripathi and H. Garg, “Machine Learning techniques for Cardiovascular Disease” in IOP Conference Series: Materials Science and Engineering, 2021, p. 012140.

[5] R. E. Ali, H. El-Kadi, S. S. Labib, and Y. I. Saad, “Prediction of potential-diabetic obese-patients using machine learning techniques” 2019.