Assessing Key Anthropometric Indicators for Non-Communicable Disease Prediction Using Machine Learning

Abstract

The most prevalent cause of global mortality is non-communicable diseases (NCDs) because over 70% of deaths in the world are attributed to NCDs; hence, the early detection and prevention of such diseases is an urgent societal issue. The research paper explores the idea of anthropometric indicators as predictive characteristics of NCD risk assessment based on machine learning (ML). The sample size was 300 adult participants (>18 years) who were recruited in Jaffna Teaching Hospital and Sabaragamuwa University in Sri Lanka, and included variables of age, gender, weight, height, 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). Normalization, outlier correction, and categorical encoding were used to preprocess the data before the training of the model. A range of supervised ML models was used, among which there were Random Forest, XGBoost, Artificial Neural Networks (ANN), Decision Tree, AdaBoost, Logistic Regression, CatBoost, and Support Vector Machine (SVM), and hyperparameter optimization was conducted with the grid search and cross-validation. Findings also indicated that ensemble learning algorithms performed better than traditional classifiers, with the highest accuracy of 98.9, XGBoost, and ANN (97.8 and 85.2), respectively, and lower accuracies of 88.5 and 85.2, respectively, with Logistic Regression and SVM. The analysis of the importance of the features indicated further that visceral fat area, body mass index, and waist-hip ratio were the most significant predictors of the NCD presence. The results indicate that ML models consuming simple anthropometric information could be used as inexpensive, scalable, and precise tools to identify NCDs at an early stage, and they could contribute to the validity of preventive healthcare strategies in resource-deprived environments with substantial potential.

Keywords: Non-Communicable Diseases, Anthropometric Indicators, Machine Learning, Random Forest, Visceral Fat Area

Introduction

Non-communicable diseases (NCDs) are the leading causes of morbidity and mortality worldwide, accounting for more than 70% of global deaths. Early detection and timely prevention are essential to reduce their burden. Machine learning (ML) has emerged as a promising tool in predictive healthcare, providing cost-effective and data-driven solutions. In particular, anthropometric indicators such as Body Mass Index (BMI), Waist-Hip Ratio

(WHR), Visceral Fat Area (VFA), and Total Body Water (TBW) have gained attention as predictors of disease risk. This study investigates the predictive capability of such anthropometric measurements in assessing NCD risk among adults in Sri Lanka, with an emphasis on identifying which features and ML models provide the most reliable performance.

Materials & Methods

The dataset consisted of anthropometric and demographic information collected from 300 adult participants, all above 18 years of age, from Jaffna Teaching Hospital and Sabaragamuwa University. The recorded variables included age, weight, height, gender, muscle body fat (MBF), total body water (TBW), percentage body fat (PBF), BMI, VFA, and WHR. Data preprocessing involved removing noise, handling missing values, normalizing features, and correcting outliers. Gender, as a categorical feature, was encoded for compatibility with machine learning models. Several supervised ML algorithms were applied, including Random Forest, Extreme Gradient Boosting (XGBoost), Artificial Neural Network (ANN), Decision Tree, AdaBoost, Logistic Regression, CatBoost, and Support Vector Machine (SVM). Model hyperparameters were optimized using grid search combined with cross-validation to ensure robust evaluation. The target variable was defined as the presence or absence of NCDs.

Results and Discussion

The results indicated that ensemble-based models outperformed traditional classifiers in predicting NCD risk. Random Forest achieved the highest accuracy of 98.9%, making it the most effective model. XGBoost and ANN followed closely, each attaining an accuracy of 97.8%. Decision Tree and AdaBoost performed moderately well, whereas Logistic Regression (88.5%) and SVM (85.2%) demonstrated comparatively lower predictive power. The findings further highlighted that anthropometric indicators such as VFA, BMI, and WHR contributed significantly to model performance, suggesting their importance in capturing risk patterns associated with NCDs. These results demonstrate that machine learning models trained on simple anthropometric data can provide accurate predictions of disease presence.

Algorithms	Accuracy (%)	Mean Squared Error [MSE] (%)	Absolute Squared Error [ASE] (%)
Random Forest	98.90	1.09	1.09
Extreme Gradient Boosting	97.80	2.19	2.19
ANN	97.80	2.19	2.19
Decision Tree	96.05	3.94	3.94
Ada Boost	93.40	6.59	6.59
Logistic Regression	88.52	11.47	11.47
Cat Boost	87.91	12.08	12.08
SVM	85.24	14.75	14.75

Conclusions (as applicable)

This study confirms that machine learning models, particularly ensemble approaches, can be highly effective for the early prediction of non-communicable diseases using anthropometric indicators. Random Forest emerged as the most accurate model, while other ensemble techniques like XGBoost and ANN also showed strong performance. Importantly, the results emphasize the predictive relevance of features such as VFA, BMI, and WHR, which can serve as reliable clinical indicators of NCD risk. By leveraging these findings, low-cost and scalable ML-based systems can be developed to support preventive healthcare, particularly in resource-limited regions such as Sri Lanka.

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.