Title: Enhancing risk prediction for diabetes, hypertension, and heart disease using SMOTE-ENN balancing with PCA and gradient boosting in healthcare AI

Author names: **Rahat** Tapon Paul , Md Assaduzzaman , Nafiz Fahad , Md Jakir Hossen

Abstract: Predicting chronic disease remains a crucial problem, particularly in low-resource environments where accurate and timely predictions are of utmost importance. Machine learning methods do not adequately generalize, encounter data imbalance and computation issues when used. In this study, an improved model that makes predictions using Synthetic minority over-sampling technique and edited nearest neighbor balancing techniques and Principal Component Analysis and Gradient Boosting to predict chronic diseases based on readily available clinical profiles is proposed. The dataset used in this study was retrieved from provide precise source: Kaggle dataset. It contains over 1500 anonymized patient records and includes 15 features such as demographic, lifestyle, and clinical measures. Standardized encoding labels, and adjusting classes using SMOTE-ENN have been completed before the PCA was conducted to improve computation speed and reduce overfitting. Decision Tree, Random Forest, LightGBM, XGBoost have been used for comparison to suggest the best performing model seen to be Gradient Boosting. PCA is performed, and the Gradient Boosting approach produces better results. Precision measures how often the classification system is correct when making a positive test result, while recall is determined using a contingency table, and the F1 score, the possibility of modeling outcomes out of 100 trials. The model proposed in the experiment provides the following outputs: Accuracy (CV: 99.33 %, CI: 98.90 %–99.50 %), Precision (CV: 99 %, CI: 98 %–99.5 %), Recall (CV: 99 %, CI: 98 %–99.5 %), F1-Score (CV: 99 %, CI: 98 %–99.5 %). The model's performance was evaluated using cross-validation, yielding an accuracy of 98.90 %. The classifying system performance is specified by the ROC-AUC ranking. It outperforms the model making indefinite projections; its ROC-AUC value is greater than 0.99. The suggested model is a robust, interpretable, and high-precision approach for the early detection of chronic conditions. Therefore, the suggested machine learning system can deliver a considerable promise with respect to creating patient-oriented outcomes.

Keywords: Chronic disease; SMOTE-ENN; PCA; Gradient boosting; Machine learning; Class imbalance; Healthcare AI

10.1016/j.ibmed.2025.100339