Diseases Classification System Using Data Mining and Machine Learning Algorithms

By:

Eng : Omnia Hosny Mohamed El sayed

Under the supervision of

Dr. Eslam Eid El Magharby

Lecturer in Information Systems Department
Faculty of Computer Science and Information Systems - Fayoum University

Prof./ Rania Ahmed Abdel Azim

Professor of digital signals , Electrical Engineering Department
Faculty of Engineering - Fayoum University

February 2024

ABSTRACT

Medical data mining is a growing field that aims to provide reliable, evidence-based medical information for doctors and researchers. Classification and data preprocessing are key strategies in this field, with the latter enhancing model quality and reliability. Diabetes, a widespread metabolic disease, affects over 422 million people globally, primarily in low or middle-income countries, causing approximately 1.5 million deaths annually. This study aims to develop a classification model that accurately identifies diabetes in patients using diagnostic data. The research study introduces a new framework that combines deep learning and standard feature selection approaches to determine if individuals have type 2 diabetes. The study evaluates the effectiveness of a deep autoencoder and conventional feature selection methods, resulting in the most optimal performance. The study uses the pima Indian diabetes dataset for benchmarking testing. The hybrid methodology's effectiveness is assessed by comparing its influence on different classification algorithms, such as neural networks, XGBoost, naïve bayes, novel K-nearest neighbor, and stacking. Statistical analysis is conducted on input features to assess their variability and significance. Shapely additive explanations is applied to highlight the most significant features obtained by the pearsonr correlation coefficient feature selection approach.
The hybrid technique achieved an accuracy rate of 83.7%, emphasizing the importance of these features in determining the final result. The proposed system has precision and recall metrics of 83.9% and 83.7%, respectively. The model's high precision and recall rates demonstrate its ability to effectively reduce false positives and false negatives, ensuring accurate diagnoses and appropriate treatment for patients.