AIN2002Report

Kerem Cantimur, Alper ..., Timuçin ...

April 2023

Introduction

Stroke is a medical condition which happens when blood supply to some region in brain is cut-off or significantly reduced. According to the World Health Organization (WHO) (reference https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html), there are near 15 million cases of stroke every year. About third of this cases results in death, other third results in permanent disability. This study aims to create an MLP model that can predict the possibility of stroke for a patient with its given status.

Related Works

There is a 2022 paper named "Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal" which collects and utilizes Electroencephalography (EEG) readings, which used Gradient Boosting algorithms. This study achieved 80% AUC with Adaptive Gradient Boosting, 77% AUC with XGBoost and 78% AUC with LightGBM. (reference https://www.mdpi.com/1424-8220/22/24/9859)
Another study in 2022 named "An Explainable Machine Learning Pipeline for Stroke Prediction on Imbalanced Data" which implemented; Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbors, Support Vector Machine and Multi-Layer Perceptron. Their LR model had the best accuracy (73.52%), specificity (73.43), and AUC (83.30) score, whereas the MLP model recorded the best sensitivity and G-Mean scores, 81.4% and 75.83%, respectively. (reference https://www.mdpi.com/2075-4418/12/10/2392)

.....

Materials and Methods

Data Description

There are two datasets, the original "Stroke Prediction Dataset" from Kaggle (reference https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset) which has 249 stroke patiens and 4861 healthy people for a total of 5110 samples. The second dataset is a synthetic one created from the original set using a deep learning model (reference https://www.kaggle.com/competitions/playground-series-s3e2/data) which has 632 stroke patients and 14672 healthy people for a total of 15304 samples. Both datasets have 11 variables: gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke; with "stroke" being the target variable.