Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220526

The Use of Machine Learning Techniques to Predict Diabetes in Patients with Cystic Fibrosis

Tajda BOGOVIČ[a], Peter KOKOL[a], Jernej ZAVRŠNIK[b] and Helena BLAŽUN VOŠNER[b]
[a] *Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia*
[b] *Community Healthcare Center dr. Adolf Drolc, 2000 Maribor, Slovenia*

Abstract. The accuracy of the prognosis of diabetes in patients with cystic fibrosis is crucial, as it highly connected with mortality and other complications. The prognosis of diabetes is a time-consuming process. Usually, it is performed by medical staff and can often lead to misdiagnosis. The aim of the study was to analyze and evaluate risk factors of developing diabetes in patients diagnosed with Cystic Fibrosis by using classification machine learning techniques. The ECFS data register was used to train and test the models. Visualization of our results using SHAP values highlights that most important features are age, antibiotic treatment, FEV1 value and lung transplant as risk predictors for diabetes.

Keywords. Machine learning, SHAP Values, Diabetes, Cystic fibrosis

1. Introduction

Cystic fibrosis (CF) is the most common recessive inherited disease among Caucasians. More than 70.000 people worldwide are diagnosed with CF. CF can cause various health problem, such as liver disease, diabetes, inflammation of the pancreas, and kidney stones [1]. Cystic fibrosis related diabetes (CFRD) is the most common complication of CF, affecting at least half of the adult population [2]. It has an impact on decreasing lung function and increasing the rate of morbidity and mortality. Early identification and treatment is crucial, as the symptoms are usually not immediately recognized. Due the lack of studies focusing on predicting risk factors for diabetes in patients with cystic fibrosis, we used different classification models for the ECFS (European Cystic Fibrosis Society) dataset to predict risk factors for this specific population.

2. Methods

ECFS dataset includes data from more than 49,000 people with CF, from 38 participating countries, and longitudinal data from 2008 to 2018. The database contains data on 389 555 entries of 55547 unique patients. The registry comprises a list of annual follow-up variables for individual CF patients that includes demographics, genetic mutations, airway colonization and microbiological infections, comorbidities and complications, transplantation, hospitalization, spirometry and therapeutic management. While machine

learning has been already used in CF [3] we developed our own optimized approach that was used on the whole ECFS dataset using the steps bellow:
Step 1: The process started from the data collection which was preprocessed.
Step 2: The k-NN approach was applied for data imputation only on features with an acceptable percentage of missing values (≤ 30%) to increase the completeness of the data.
Step 3: Synthetic Minority Oversampling Technique was used to balance the data.
Step 4: A total of 43 attributes were divided into training and test data set in the ration 70:30 and importance of features were explained using SHAP (Shapely Additive Explanations) values to make results easier to understand for medical professionals [4].

3. Results and Discussion

Table 1, presents the Accuracy and Area under ROC curve (AUC) (for three popular machine learning models and Figure 1 the SHAP values of the best model (Cat Boost).

Table 1. Models' accuracy and AUC

Model	Accuracy [%]	AUC [%]
Catboost Model	91±0.06	92±0.01
Decision Tree Classifier	81±0.01	81±0.02
Random Forest Classifier	84±0.02	83±0.03

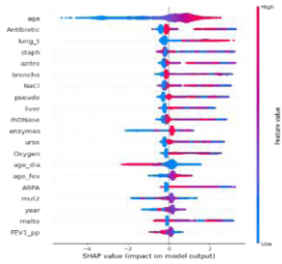

Figure 1. Cat Boost Model - SHAP values

Cat Boost classification indicates that most important features in patients with both diabetes and CF are age, antibiotic treatment, and lung transplant, as well as FEV1 (Volume of air blown). Same attributes were also most prolific in other models. Our study also highlights the advantage of presenting the results of machine learning with SHAP values which can be also easily understood by non-machine learning experts in our case physicians and other health professionals.
Funding: The research was partially funded by EC H2020 Project STAMINA- GA 883441

References

[1] De Boeck K. Cystic fibrosis in the year 2020: A disease with a new face. Acta paediatrica. 2020 May;109(5):893-9. doi: 10.1111/apa.15155.
[2] Iafusco F, Maione G, Rosanio FM, Mozzillo E, Franzese A, Tinto N. Cystic Fibrosis-Related Diabetes (CFRD): Overview of Associated Genetic Factors. Diagnostics. 2021 Mar;11(3):572.
[3] Alaa AM, van der Schaar M. Prognostication and Risk Factors for Cystic Fibrosis via Automated Machine Learning. Sci Rep. 2018 Jul 26;8(1):11242.
[4] Lu S, et al. Understanding Heart Failure Patients EHR Clinical Features via SHAP Interpretation of Tree-Based Machine Learning Model Predictions. In AMIA An Symp Proceedings 2021 (Vol. 2021, p. 813).