J. Mantas et al. (Eds.)

doi:10.3233/SHTI230459

A Conceptual Framework to Predict Disease Progressions in Patients with Chronic Kidney Disease, Using Machine Learning and Process Mining

Nichalini KANDASAMY[a], Thierry CHAUSSALET[a,1] and Artie BASUKOSKI[a]
[a] *School of Computer Science & Engineering, University of Westminster, London, UK*

Abstract. Process Mining is a technique looking into the analysis and mining of existing process flow. On the other hand, Machine Learning is a data science field and a sub-branch of Artificial Intelligence with the main purpose of replicating human behavior through algorithms. The separate application of Process Mining and Machine Learning for healthcare purposes has been widely explored with a various number of published works discussing their use. However, the simultaneous application of Process Mining and Machine Learning algorithms is still a growing field with ongoing studies on its application. This paper proposes a feasible framework where Process Mining and Machine Learning can be used in combination within the healthcare environment.

Keywords. Process Mining, GMM algorithm, Hybrid algorithm, DREAM, Chronic Kidney Disease

1. Introduction

The study will focus on Chronic Kidney Disease, which 3.5 million in the UK alone.[1] CKD is comorbid with diabetes, hypertension and cardiovascular disease.[2] The disease progresses throughout 5 stages and at various rate depending on the patient's other medical conditions. With the progression of CKD, patients are at higher risk of developing cardiovascular disease and lower quality of life. An accurate and advanced evaluation of the disease progression can help clinicians and patients to get the most beneficial treatment to slow down the progression of the disease. Early knowledge of possible future diagnoses can guide medical practitioners in administrating an appropriate medical examination and treatment.

When treating a patient, their medical details are recorded and stored into a database which can be accessed by clinicians to view and analyze patient records. Those patient records are known as Electronic Health Records (EHR) and have historical medical information about patients that are regularly maintained and updated by the database owner (e.g., hospital, GP practices). Electronic Health Records are formed of various

[1] Corresponding Author: Thierry Chaussalet, E-mail: chausst@westminster.ac.uk

datasets that contain laboratory results, radiographies, clinical notes and observations, progress reports, historical medication records and patients' personal information. The principal purpose of the EHR is to accurately record patients' information and medical progression to avoid duplication and inappropriate administration of treatment. [3] Each historical diagnosis is captured with a timestamp and diagnosis code (*ICD codes – International Classification of Diseases*) that can be mapped onto an event log. The event log can then be mined to understand the existing treatment flow and gain an understanding of the disease progression using historical health records. The process model can help convert timestamp events into a variable that can be used as input variable for the prediction model and determine the next probable diagnosis code. Existing studies mainly focus on either the sole application of process mining or machine learning to predict disease progression. The use of timestamped variables is never considered due to the complexity of engineering those variables for machine learning models.

In this study, the MIMIC-IV database, also known as Medical Information Mart for Intensive Care, will be used to extract event log information and proceed with disease progression's prediction.[4] The database contains anonymized data of 40,000 patients' who were admitted at the Beth Israel Deaconess Medical Centre's ICU. The database is structured in 2 modules, '*hosp'* and '*icu'*, where each of them contains historical medical records and associated events within the hospital and ICU.

We propose a framework that combines the application of process mining and machine learning to efficiently predict chronic kidney disease progression. A major objective of the study is to understand the impact of using timestamped information as part of the prediction process and how well it improves the framework's performance. The development of the framework will be formed of various steps through data Clustering, Process Mining and Predictive analytics using Hybrid Radial Basis Function Neural Network (RBF-NN). The engineering of the time-stamped variables will be completed using the DREAM algorithm. The integrated model's output can help medical teams in providing precocious treatment to prevent the progression of the disease or to reduce further complications.

2. Existing framework

At present, only a small number of studies have been conducted in regard to the application of process mining and machine learning in a healthcare environment, albeit for the purpose of predicting survival rate rather than disease progression. As such, Theis et al. [5], established a framework using process mining and deep learning algorithms to improve in-hospital mortality prediction for diabetes patients admitted to ICU. The objective of the study is to improve the existing severity scoring system by incorporating patients' historical hospital admissions data as the existing scoring system only analyzed patients' current clinical information. The development of the framework was completed following 3 core stages: Process Discovery, then DREAM (Decay REplAy Mining) algorithm, Long Short-Term Memory (LSTM) Neural Network modelling and Evaluation. The developed methodology outperformed the existing scoring methodology, with an AUROC score of 0.873 against 0.713 for Linear Regression and 0.709 for Random Forest.

Another combined application of Process Mining and Deep Learning was scrutinized by Pishgar et al. [6] to predict the survival of hospitalized Covid-19 patients. Numerous studies focused on the application of machine learning algorithms to predict outcomes of hospitalized Covid-19 patients. The study conducted by Pishgar et al. aims to develop a framework that will predict the survival rate of patients at an interval of 6h within the first 72h of being admitted, by using time as a variable in the modelling part. As an initial step, a process model was built to discover the existing patients' trajectories. The output of the process model was fed into a Neural Network model after being processed by the DREAM algorithm which parameterizes time based on the trajectory. The final model was evaluated to an AUROC score of 0.80, which outperforms singular machine learning models that were tested.

3. Methodology

Based on the frameworks that have already been explored and discussed, the proposed framework contains an unsupervised and supervised machine learning approach around process mining. The framework performs clustering using Gaussian Mixture Model (GMM) prior to process discovery, then applies Hybrid Radial Basis Function Neural Network as predictive algorithm to output probable diagnoses.

As seen with the study conducted by Pishgar et al. [6], the data was highly imbalanced due to mortality and comorbidities being more prevalent in one demographic group than others. To deal with imbalance datasets and demographic factors that influence the process, a data clustering approach is proposed to cluster the event logs based on patients' information. This would allow us to run the process mining in parallel across various groups. Gaussian Mixture Model is an unsupervised algorithm, with the capability of clustering data using probabilistic methods. An example application of GMM for healthcare purposes was conducted by Abbi et al. [7], who analyzed the application of GMM to group patients based on their length of stay in hospital. The result demonstrated that GMM successfully reflected the clinical pattern visible across stroke patients.

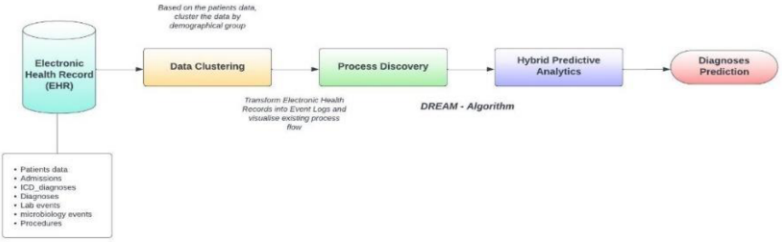
Based on the patients data, cluster the data by demographical group
Electronic Health Record (EHR)
Data Clustering
Process Discovery
Hybrid Predictive Analytics
Diagnoses Prediction
Transform Electronic Health Records into Event Logs and visualise existing process flow
DREAM - Algorithm
• Patients data
• Admissions
• ICD_diagnoses
• Diagnoses
• Lab events
• microbiology events
• Procedures

Fig. 1 : Proposed Process Mining / Hybrid Machine Learning framework to predict diagnoses in patients with chronic illness.

Figure 1. Proposed Process Mining – Machine Learning framework to predict diagnoses in patients with chronic illness.

As shown in Figure 1, upon clustering the event logs, a process model must be generated using process discovery algorithms to understand the existing disease trajectories. The output of the process model is then processed using the DREAM algorithm, which is available in ProM, an open source framework that supports a wide range of process mining techniques [8]. The DREAM algorithm allows the generation of Timed State Samples (TSS) which are fractions of the process model for each patient with information about their health condition up to the predicted time.[5]

That information allows us to feed the Hybrid Predictive Analytics component with time defined variables and have historical medical information about a patient's condition. The reviewed frameworks focused on the application of a Deep Learning algorithm for prediction. However, although deep learning models can be high performing and fast, they can overfit and require high computational power. To tackle this issue, we are using hybrid machine learning models, which are a combination of two algorithms that may not perform efficiently while being utilized independently.

Therefore our Hybrid Predictive Analytics component will combine a Tree-based algorithm (e.g., XGBoost, Random Forest, AdaBoost) with RBF-NN, which has been widely used on its own for disease prediction and classification. The optimal tree-based model will be identified after modelling various tree-based models and evaluating their performances. Bayesian optimization will be used as an optimizer to fine tune the parameters of the hybrid model. The model evaluation will be conducted using several metrics: AUROC, precision, recall and accuracy.

4. Future Work

This paper discussed the various aspects of establishing an integrated process mining and machine learning framework to predict disease progression for patients with chronic kidney disease. As a next step in the implementation of the framework, the discussed methodology will be tested and validated on the MIMIC-IV's EHR data to ensure the feasibility of the framework. Various data pre-processing steps will be taken to only focus on chronic kidney disease and its associated comorbidities. The final output of the model will be compared against a singular process model's output and a machine learning model's output to analyze the performance of the integrated model. The framework must be flexible and reproducible to ensure accurate results while being applied to various clinical data.

References

[1] Kidney Care UK, Facts about kidneys [Internet]. Available from : https://www.kidneycareuk.org/news-and-campaigns/facts-and stats/#:~:text=Kidney%20disease-,Around%203.5%20million%20people%20in%20the%20UK%20have%20Chronic%20Kidney,the%20biggest%20causes%20of%20CKD.

[2] C. MacRae, SW.Mercer, B. Guthrie, Comorbidity in chronic kidney disease: a large cross-sectional study of prevalence in Scottish primary care, Feb 2021, 25;71(704):e243-e249. DOI: 10.3399/bjgp20X714125. PMID: 33558333; PMCID: PMC7888754.

[3] R.S.Evans, Electronic Health Records: Then, Now and in the Future, Yearbook of Medical Informatics, May 2016, DOI: 10.15265/IYS-2016-s006

[4] MIMIC-IV Online Documentation, MIMIC-IV [Internet], Available from : https://physionet.org/content/mimiciv/2.2/

[5] J. Theis, W.L.Galanter, A.D.Boyd, H.Darabi, Improving the In-Hospital Mortality Prediction of Diabetes ICU Patients using a Process Mining/Deep Learning Architecture, Jan 2022, IEEE, Vol.26, No 1.

[6] M. Pishgar, S.Harford, J.Theis, W.Galanter, A process mining-deep learning approach to predict survival in a cohort of hospitalized COVID-19 patients, BMC Med Inform and Decision Making, 22, 194, July 2022, https://doi.org/10.1186/s12911-022-01934-2

[7] R.Abbi, E. El-Darzi, C. Vasilakis, P.H. Millard, A Gaussian mixture model approach to grouping patients according to their hospital length of stay,2008, Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS '08) IEEE . pp. 524-529

[8] ProM Tools [Internet]. Available from: https://promtools.org/