Classification of disease severity using machine learning algorithms: An analysis for chronic kidney disease in the US

Verma V, Rastogi M, **Khan S**, Bharti S, Pandey S, Sanyal S, Bansal V, Gaur A, Daral S, Kukreja I, Nayyar A, Roy A

RWD185

Objectives

Chronic kidney disease (CKD) is a growing concern in the US, with approximately 14% of the adults having kidney ailments. Since lack of proper healthcare facilities and inadequate awareness lead to a progression of the disease, it is of ardent need to identify the level of kidney dysfunction and undertake prompt treatment actions.

In this study, we evaluate machine learning (ML) algorithms to predict the classification of CKD severity for appropriate and timely intervention to mitigate the disease progression.

Methods

- Optum® de-identified and consolidated Electronic Health Records (EHR) data of Market Clarity database was used for this study.
- The analysis was based on the study period from 2015 till 2022, with a look-back period from 2015 to 2019 and an identification period from 2021 to 2022. A 1-year window (2020-2021) between the baseline and the identification period was considered , which masked the patients' medical history for that year.
- Diagnosis of CKD was recognized as the index event, which led to a total of 23,840 CKD patients, aged 45 years and older for the study.
- The severity of CKD was considered as the outcome of interest, with multi-level classifications of the severity stages calculated based on ICD-10-CM codes of CKD. The following are the representations of the CKD stages used:
    - Mild CKD - Severity stages 1, 2
    - Moderate CKD- Severity stages 3, 4
    - Severe CKD- Severity stage 5, End Stage Renal Disease
- Patients' demographic characteristics, underlying comorbidities and laboratory tests were considered, encompassing 30 predictors.
- Variance Inflation Factor (VIF) had been used to remove multicollinearity.
- Supervised ML techniques which include Logistic regression, Decision tree and Random Forest were used to anticipate the disease stages based on the prevailing symptoms.
- Synthetic Minority Over-sampling Technique (SMOTE) had been used for handling data imbalance issue.

Results

Figure 1: Feature importance of top 10 predictors

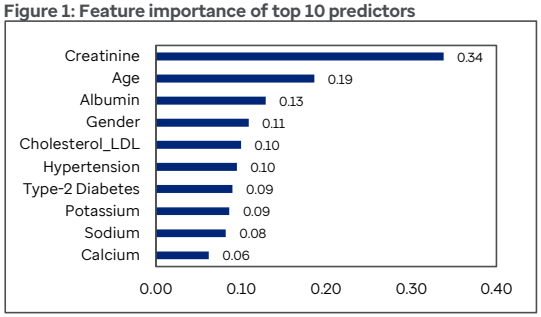
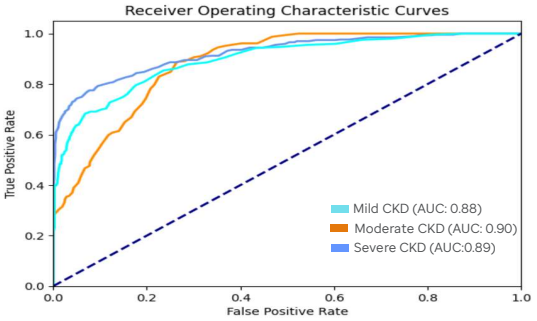


Table 1: Performance of Random Forest model

Model	Sensitivity	Precision	F1-Score	ROC-AUC
Mild CKD	0.84	0.84	0.84	0.88
Moderate CKD	0.85	0.84	0.85	0.90
Severe CKD	0.85	0.86	0.86	0.89

ROC-AUC: Receiver operating characteristics- Area under curve

Figure 2: Receiver operating characteristics (ROC) curve for



- The accuracy of disease severity was evaluated for each of the models to obtain the most precise results. Random forest provided the best accuracy of 85% .
- Among the predictors used for the overall analysis, top 10 predictors were selected based on feature importance (Figure 1). Laboratory tests like creatine and albumin were found to have high feature importance score of 0.34 and 0.13 respectively. Demographic characteristics also exhibited comparatively high importance, which includes age (0.19) and gender (0.11). Hypertension (0.10) and Type-2 Diabetes (0.09) were the two comorbidities which ranked among the top 10 predictors.
- High sensitivity, precision, F1 and ROC-AUC score were seen for all stages of CKD (Table 1).
- The AUC of each CKD stage was calculated using the predicted probability from Random Forest algorithm. The AUC score for mild, moderate and severe CKD were 0.88, 0.90 and 0.89 respectively (Figure 2).

Conclusions

- The vast repository of clinical data in the EHR database is effectively utilized in our study for an accurate prediction of CKD progression.
- This study shows the feasibility of ML techniques in evaluating the prognosis of CKD based on easily accessible features.
- This model can be implemented as a decision support tool to minimize prescription errors and assist the providers with timely and accurate prognosis.
- Hospital visits and readmissions can also be reduced to a notable extent by properly utilizing the exhaustive medical data, enabling an efficient patient care delivery.

Optum