MACHINE LEARNING PREDICTION MODELS FOR CHRONIC KIDNEY DISEASE

N. SATYANADAM[1], CHEVULA SINDHU[2], TEJAVATH SRAVANTHI[3], SIRISALA SUJATH[4]

ASSOSCIATE PROFESSOR[1], UG SCHOLAR [2, 3 &4]

DEPARTMENT OF CSE, BHOJ REDDY ENGINEERING COLLEGE FOR WOMEN, VINAY NAGAR, HYDERABAD-59

ABSTRACT:

Chronic kidney disease (CKD) represents a heavy burden on the healthcare system because of the increasing number of patients, high risk of progression to end-stage renal disease, and poor prognosis of morbidity and mortality. The aim of this study is to develop a machine-learning model that uses the comorbidity and medication data obtained from Taiwan's National Health Insurance Research Database to forecast the occurrence of CKD within the next 6 or 12 months before its onset, and hence its prevalence in the population. A total of 18,000 people with CKD and 72,000 people without CKD diagnosis were selected using propensity score matching. Their demographic, medication and comorbidity data from their respective two-year observation period were used to build a predictive model. Among the approaches investigated, the Convolutional Neural Networks (CNN) model performed best with a test set AUROC of 0.957 and 0.954 for the 6-month and 12-month predictions, respectively. The most prominent predictors in the tree-based models were identified, including diabetes mellitus, age, gout, and medications such as sulfonamides and angiotensins. The model proposed in this study could be a useful tool for policymakers in predicting the trends of CKD in the population. The models can allow close monitoring of people at risk, early detection of CKD, better allocation of resources, and patient-centric management.

INTRODUCTION Chronic Kidney Disease (CKD) is a condition resulting in insufficient kidney function, where patients have to live with a compromised quality of life. Asia has the highest prevalence of CKD in the world, led by Japan and followed by Taiwan. In Taiwan, CKD has been the eighth leading cause of death since 1997. Compared to other countries, Taiwan has higher incidences and mortality rates, with the prevalence increasing from 1.99% in 1996 to 9.83% in 2003 [1], while awareness about CKD has remained low [2]. CKD is a substantial financial burden on patients, healthcare services, and the government. Treatments of the ESRD with Renal Replacement Therapy are either expensive (hemodialysis and peritoneal dialysis) or complex (transplantation). Taiwan has about 0.1%–0.2% of the population receiving dialysis—contributing to about 7% of the total budget of the National Health Insurance (NHI) program [3].

The association of CKD with other chronic diseases also exacerbates the situation. From the public health perspective, it is therefore imperative to be able to predict the trends in terms of CKD prevalence so iction models can be developed [5,6]. At the patient level, a physician can assess the onset of CKD using laboratory tests by looking at standard parameters such as the glomerular filtration rate (eGFR) and the albumincreatinine ratio [7]. On the other hand, from the public health perspective, laboratory data is typically not available on a large scale. However, two types of data can generally be extracted from the insurance companies' databases: diagnoses and medications for each patient's visit at the hospital. Common approaches for developing disease prediction models with EHR data involve collecting clinical and laboratory data from sources such as billing or claims data, discharge summaries, patient history, etc., and building models on features extracted from them. Some previous studies attempted to use longitudinal data to capture temporal patterns to develop disease prediction models for CKD. Ren et al. (2019) developed a predictive model for kidney disease among patients with hypertension from EHR consisting of textual and numeric information. They proposed a neural network framework based on Bidirectional long short-term memory and auto-encoders to encode the textual and numerical information, respectively. They performed under-sampling to balance the data. They achieved 89.7% accuracy with 10-fold cross-validation [8]. Song et al. (2020) presented a one-year prediction model using a landmark-boosting approach based on gradient boosting machines for CKD among diabetes patients with an AUROC of 0.83. They analyzed longitudinal data containing several clinical observations derived from EHR and billing data [9]. Fenglong et al. (2018) proposed a general framework using posterior regularisation techniques that incorporate prior medical information from the EHR for prediction models. The constraint feature design in the framework took into consideration patient characteristics, underlying disease, disease duration, genetics, and family history. The patient characteristics included sex, age, and ethnicity. While using prediction models for a certain disease, the framework took into consideration the diagnosis of the underlying disease that would be related to the occurrence of the main disease to be predicted [10]. Another similar work by Katsuki et al. (2018) predicted the risk of entering the second stage diabetic nephropathy from the first stage using EHR data consisting of sequences of lab test results. They used convolutional autoencoders to encode the temporal features and achieve performance better than baseline models [11]. Similarly, some studies used non-temporal EHR data to develop disease prediction models. Song et al. (2019)

extracted several significant clinical features from EHR data using an ensemble feature selection method to predict the risk of kidney disease among diabetes patients. They achieved an AUROC of 0.71 on an external validation set [12]. Jardin et al. (2012) predicted kidney-related outcomes among diabetes patients using the Cox proportional hazard regression on the ADVANCE cohort, which comprises demographic, behavioral, and physical information, and relevant lab values. They achieved a C statistics of 0.847 on predicting major renal events [13]. Dovgan et al. (2020) predicted the onset of renal replacement therapy three, six, and 12 months after the CKD diagnosis. For their 12-month prediction, they achieved an AUC of 0.773 In this paper, we aimed to develop machine-learning models that predict the onset of CKD within the next 6 and 12 months. The model is based on the insurance claims data (age, sex, comorbidities, and medication) over an observation period of 24 months. Further, we aim to assess the reliability of the models by identifying the comorbidities and medications that impact the development of CKD

EXISTING SYSTEM

- Corinne Isnard Bagnis, Jack Edward Heron, David M. Gracey et al. [1] conducted a report on Chronic Kidney Disease and its connection to more deplorable outcomes It shows that controlling blood pressure with angiotensin converting enzyme inhibitors and angiotensin receptor blockers slows the progression of CKD in HIV patients, particularly when proteinuria is present. Y. Liu, J. Qin, C. Feng, L. Chen, C. Liu, and B. Chen et al. [2] reveals that data imputation and sample diagnosis are possible with CKD. The integrated model presented in this paper can achieve sufficient accuracy using the KNN algorithm. Since the dataset contains two classes, Chronicle Kidney Infection and Not Chronic Kidney Disease, the model cannot investigate the stages of chronic kidney disease. A. S. Anwar and E. H.
- A. Rady et al. [3] uses lab dataset of 361 persistent kidney sickness patients . It uses PNN, SVM, and MLP algorithms to calculate period of chronic kidney sickness. This examination suggests that theprobabilistic neural organization calculation is best performing calculation that can be utilized by doctors to kill demonstrative and treatment mistakes. M. N. Amin, A. Al Imran and F. T.

- Johora et al. [4] analyze model performance on real (imbalanced) data and model performance on oversampled (balanced) data using logistic regression and feed forward neural networks. Feed forward neural networks showed the best results for both real and oversampled data, with 0.99 Recall, 0.97 Precision, 0.99 F1-Score and 0.99 AUC score.
- K. S. Vaisla, N. Chetty and S. D. Sudarsan et al. [5] recommended On the CKD dataset, attribute assessment and classification models were used. The attribute evaluator model performed better by decreasing the number of attributes from 25 to 6, 12, and 7. P. Arulanthu and E. Perumal et al. [6] utilizes JRip, SMO, Naive Bayes, algorithms and analyses that JRip generate best performance.
- P. Manickam, K. Shankar, M. Ilayaraja and G. Devika et al. [7] uses Ant Lion Optimization (ALO) technique to choose ideal features for classification. This optimization results in better classification accuracy for deep neural network. R. Shinde, Maurya, R. Wable,S. John, R. Dakshayani and R. Jadhav, et al. [8] To slow the progression of CKD and to follow the recommended diet plans, use the potassium zone, which is computed using blood potassium levels. R. Yadav and S. C. Jat et al. [9] investigate the relation of various methods of selection and dimensionality reduction to the performance of chronic disease classification and prediction.

DISADVANTAGES

- The system doesn’t have a method to find CKD stage identification.
- The Multilayered Perceptron (MLP) separator was not used to predict HBV-induced hepatic cirrhosis, and the findings indicate that the MLP separator provides excellent predictive results for liver disease, particularly in HBV-related patients with liver failure.

PROPOSED SYSTEM

There are several machine learning algorithms used in literature for CKD classification. In this paper, we have built 6 ML models using, KNN,SVM, random forest, decision tree, ada-boost and xg-boost algorithms, along with a simple deep neural network to classify weather a patient has CKD or not. The flow of the proposed experimental setup is depicted in Figure 2For binary classification situations, A SVM (support vector machine) is a classification-based supervised machine learning model. K-nearest neighbors (KNN) algorithm utilizes feature comparing to

predict a value according on how closely it is similar in the training dataset. A decision tree is used to visually represent decisions of classification. Often, a single decision tree is not sufficient for producing effective classification accuracy. Random Forest algorithm solves this problem by leveraging multiple decision trees. AdaBoost algorithm, also called adaptive boosting, is a boosting technique used as an ensemble method in machine learning. It aims to convert a set of weak classifiers into a strong one by reassigning the weights to each instance. XG Boost (eXtreme Gradient Boosting) is another boosting algorithm that uses a gradient boosting framework. Other than machine learning, many researcher have utilized feature based deep neural network (DNN) for better classification results. Deep neural networks are capable of detecting crucial disease since they use several layers of nodes to accomplish high-level functions from input data. Before applying classification algorithm, we have eliminated few features using feature selection method.

ADVANTAGES

- RFE, or Recursive Feature Elimination, is a widespread attributes selection algorithm that selects the features (columns) in a training dataset that are more or more important in predicting the target variable.
- Automated computer aided diagnose for CKD is a process of getting stage information using patient data such as age, blood pressure, blood test reports. Yu et al. [2] has utilized the Support Vector Machine (SVM) algorithm to recognize and anticipate diabetic and pre-diabetic patients.

Modules

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Train and Test Data Sets,View Trained and Tested Accuracy in Bar Chart,View Trained and Tested Accuracy Results,Find Kidney Disease Status,Find Kidney Disease Ratio,View All Kidney Disease Positive Details,Download Trained Data Sets,View Kidney Disease Ratio Results,View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like POST DIABETIC DATA SETS,VIEW YOUR PROFILE.

CONCLUSION

In this study, we developed and evaluated a series of artificial intelligence-based models considering minimum variables such as sex, age, comorbidities, and medications. These models predict patients' risk of developing chronic kidney disease after a period of 6 or 12 months. Among various models tested, convolutional neural networks (CNN) performed best, with an AUROC metric of 0.957 and 0.954 for 6 and 12 months, respectively. To see which features are the most prominent for prediction, we looked at the tree-based LightGBM model. The most prominent features included diabetes mellitus, age, gout, and use of sulfonamides and angiotensins, which are all reasonable in view of CKD. From a policymaker's point of view, these ML-based models could be efficiently used in resource management and initiating public health initiatives such as closely monitoring and early detection of CKD. Clearly, for the application of such models into clinical practice dealing with individual patients, the feature set would have to be expanded to include laboratory measurements and possibly lifestyle information, which falls within the scope of future work.

REFERENCES

1. Kuo, H.-W.; Tsai, S.-S.; Tiao, M.-M.; Yang, C.-Y. Epidemiological Features of CKD in Taiwan. Am. J. Kidney Dis. 2007, 49, 46–55. [CrossRef] [PubMed]

2. Hsu, C.-C.; Hwang, S.-J.; Wen, C.-P.; Chang, H.-Y.; Chen, T.; Shiu, R.-S.; Horng, S.-S.; Chang, Y.-K.; Yang, W.-C. High Prevalence and Low Awareness of CKD in Taiwan: A Study

on the Relationship Between Serum Creatinine and Awareness from a Nationally Representative Survey. Am. J. Kidney Dis. 2006, 48, 727–738. [CrossRef] [PubMed]

3. Navva, P.K.R.; Sreepada, S.V.; Nayak, K.S. Present Status of Renal Replacement Therapy in Asian Countries. Blood Purif. 2015, 40, 280–287. [CrossRef] [PubMed]

4. Vanholder, R.; Annemans, L.; Brown, E.; Gansevoort, R.; Gout-Zwart, J.J.; Lameire, N.; Morton, R.L.; Oberbauer, R.; Postma, M.J.; Tobelli, M.; et al. Reducing the costs of chronic kidney disease while delivering quality health care: A call to action. Nat. Rev. Nephrol. 2017, 13, 393–409. [CrossRef] [PubMed]

5. Callahan, A.; Shah, N.H. Machine Learning in Healthcare. In Key Advances in Clinical Informatics; Elsevier: Amsterdam, The Netherlands, 2017; pp. 279–291. [CrossRef]

6. Adkins, D.E. Machine Learning and Electronic Health Records: A Paradigm Shift. Am. J. Psychiatry 2017, 174, 93–94. [CrossRef] [PubMed]

7. Collins, A.J.; Vassalotti, J.A.; Wang, C.; Li, S.; Gilbertson, D.T.; Liu, J.; Foley, R.N.; Chen, S.-C.; Arneson, T.J. Who Should Be Targeted for CKD Screening? Impact of Diabetes, Hypertension, and Cardiovascular Disease. Am. J. Kidney Dis. 2009, 53, S71–S77. [CrossRef] [PubMed]

8. Ren, Y.; Fei, H.; Liang, X.; Ji, D.; Cheng, M. A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records. BMC Med. Inform. Decis. Mak. 2019, 19, 131–138. [CrossRef] [PubMed]

9. Johansson, M.; Buijs, J.O.D.; Song, X.; Waitman, L.R.; Yu, A.S.; Robbins, D.C.; Hu, Y.; Liu, M. Longitudinal Risk Prediction of Chronic Kidney Disease in Diabetic Patients Using a Temporal-Enhanced Gradient Boosting Machine: Retrospective Cohort Study. JMIR Med. Inform. 2020, 8, e15510.