International Journal of Advance Research in Science and Engineering
Volume No. 14, Issue No. 05, May 2025
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Chronic kidney disease prediction using machine learning with Web application

Prathik V Shenoy[1], Siva Selvan[2]

[1]CSE Department, Manipal Institute of Technology, Manipal Academy of Higher Education
Manipal, Karnataka, India
[2]Assistant Professor (Senior scale), CSE Department,
Manipal Institute of Technology Manipal Academy of Higher Education
Manipal, Karnataka, India

ABSTRACT

The integration of computer vision and machine learning has brought about significant changes, especially in healthcare. This paper explores the substantial influence of machine learning on current healthcare practices, highlighting its role in efficient data analysis and decision-making. With machine learning algorithms, healthcare providers can now im prove disease diagnosis and treatment by quickly and accurately interpreting large datasets. These advancements in machine learning have made advanced diagnostic tools more accessible, reducing healthcare disparities and promoting more equitable patient care.

I. INTRODUCTION

Chronic Kidney Disease (CKD) is a global health concern, impacting millions and placing a significant economic strain on healthcare systems [1]. Early identification and prediction of CKD are crucial for prompt intervention and effective management. Machine learning, known for its capability to analyze extensive datasets and identify patterns, has proven to be a valuable tool for CKD prediction

Despite its potential, leveraging machine learning requires careful data curation and significant computational resources for model training [2]. This paper underscores the importance of supervised machine learning, a method that can predict future events based on past labeled examples. It involves exam ining known training datasets to develop an inferred function, which then facilitates predictions about output values.

As we examine the integration of machine learning and computer vision in healthcare, we also consider the challenges related to scalability, data quality, and model interoperability [3]. Additionally, we offer insights into the promising future of this interdisciplinary field, anticipating more efficient and personalized healthcare solutions. This paper enhances the understanding of the current landscape and paves the way for future advancements in healthcare driven by machine learning powered computer vision.

II. RELATED WORK

• In a study by Dhivya S. and her colleagues, they analyzed a Chronic Kidney Disease (CKD) dataset which they got from the UCI repository. This dataset included records from 400 patients, featuring 26 attributes relevant to CKD nosis [2]. Their main goal was to preprocess this dataset for effective use in machine learning applications.

To improve the dataset's suitability for analysis, the researchers handled missing data by substituting it with a specific value (0) [4]. They then applied various machine learning techniques to transform the dataset for further examination. After preprocessing, the authors employed multiple machine learning algorithms to pinpoint the most critical factors influencing CKD diagnosis. Their analysis identified five key features that significantly impacted diagnostic outcomes [5]. These insights were then applied to the entire dataset, including its columns, rows, and individual data points."

• According to Shengguo Hu et al. (2009) , a significant area of focus and challenge in machine learning lies in the classification of algorithms. several classification methods have been studied and proven effective in prac tical applications. However, a major concern arises when these methods are applied to imbalanced datasets [6], particularly affecting the performance of the minority class. In practical scenarios such as detecting fraudu lent exchanges, network intrusion detection and medi cal diagnostics (e.g., disease identification), imbalanced datasets are common. Unfortunately, existing classifica tion methods often perform suboptimally when dealing with imbalanced datasets. This imbalance issue is a critical consideration in various real-world applications, underscoring the need for enhanced methodologies to effectively address such scenarios..

• Extreme Gradient Boosting (XGB) is a highly efficient implementation of gradient boosting that focuses on identifying the optimal tree model. XGB is distinguished by its use of second-order gradients, which provide additional insights into the gradient direction to reduce the loss function effectively. Unlike a basic model like a decision tree, which minimizes the overall model cost using the loss function as a proxy, XGB [3] uses the second-order derivative to provide output. To improve the model's generalization ability, XGB incorporates ad vanced L1 and L2 regularization techniques. These tech niques help refine the model's performance and ensure it can adapt effectively to various datasets.

• Hussianzadah provided a diagnostic approach using 4 dif ferent classifiers: Decision Tree, Support Vector Machine (SVM), Multi-Layer Perception (MLP), and Naive Bayes. They applied these to 3 distinct datasets, each having 14, 12, and 13 features, respectively [7]. The SVM classifier basiaclly achieved an accuracy of 91%. Additionally, an design for diagnosing kidney disorders using SVM was presented [8], incorporating feature selection methods with the use of wrapper and filter techniques on the dataset. The highest accuracy was found with SVM and a filtered subset evaluator using the BFS traversal feature selection method, achieving an accuracy value of 98.5

• In a previous study [9], a single Chronic Kidney Disease (CKD) attribute was derived from year-long temporal data obtained from electronic health records (EHRs). But , this approach had a limitation: it excluded new patients without EHRs from benefiting. Notably, these studies utilized black-box classification methods for constructing models, presenting outcomes and selected attributes. The primary concern raised was the lack of interpretability in the decision-making process of these diagnostic mod els, which could lead to potentially adverse, even life threatening, consequences. Another challeng identified was fewer number of appropriate criteria for choosing specific attributes for decision making of the model [7]. Therefore, there is a recognized need for ongoing research to develop interpretable machine learning (ML) models for computer-aided diagnostic systems. The aim is to enable clinicians to more effectively evaluate model decisions and understand the role of individual model attributes in the decision-making process. This research seeks to address the shortcomings of opaque model architectures and improve the safety and effectiveness of diagnostic systems.

• Md Ashique Islam et al. (2020) [10] demonstrated that the Naive Bayes classification is an effective approach. Simple classifiers based on probability are the main feature of the Naive Bayes classifier. However, in our study, logistic regression outperformed the Naive Bayes classifier in predicting CKD. While Naive Bayes achieved an accuracy of 93.9056%, this was less than the accuracy of logistic regression.

A. Materials and methods

The dataset that has been referred to in this study was extracted from a web page named as Kaggle and was collected through a hospital in India in July 2015. It comprises 400 samples, with 250 named as chronic kidney disease and 150 as non-chronic kidney disease [4]. This dataset includes 33 predictive features, such as blood pressure, specific gravity, albumin, sugar, red blood cell, blood urea, serum creatinine, sodium, potassium, hemoglobin, white blood cell count, red blood cell count, and hypertension. Out of these 33 features, only 24 were considered.

For data preprocessing, the response data was evenly dis tributed to ensure reliable results. The data was randomly split

to create a balanced dataset, with 340 samples used for training and 60 for testing using a 15% holdout method. Various classification techniques were evaluated, including Support Vector Machines (SVMs) [8], k-nearest neighbors (kNN), and logistic regression. The highest accuracy was possible with Gaussian SVM and logistic regression using 10-fold cross-validation, where the data was seperated into 10 parts, each containing 90% of the original data. To maximize accuracy, seven training iterations were conducted. Each iteration showed some performance variation due to stochastic elements inherent in machine learning algorithms, where slightly different models are learned from the same data [8]..

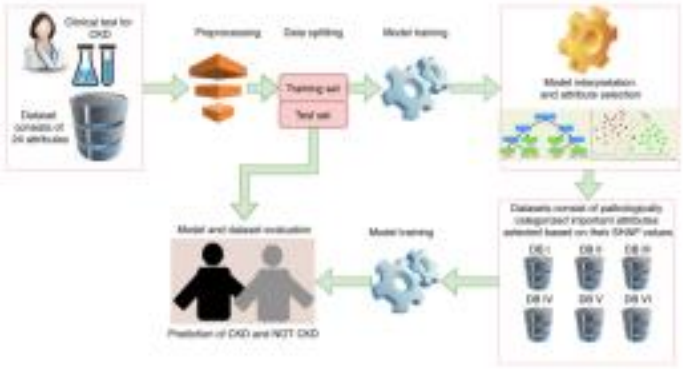

Fig. 1. model functioning

B. pre processing

• pre-processing of the unrefined data has been done by tak ing out the missing values or elements to upgrade predic tion capabilities. We have also done data-transformation [6]so that it is useful for the ML models, which are constrained to process non-numerical data. This kind of data or values which are presented in the dataset are in the form of ‘present’, ‘not present’, ‘normal’, ‘abnormal’, ‘yes’, ‘no’, ‘good’, and ‘poor’.

• The non-numerical data are identified and transformed into numbers. The ‘normal’, ‘present’, ‘yes’, and ‘good’ values for nominal or selected attributes are replaced or modified by ‘1’ and ‘abnormal‘, ‘notpresent’, ‘no’, and ‘poor’ values are replaced by ‘0’.Missing values are associated to real-world data.Ignoring of the iteration that has the missing elements can be taken as the simplest means of a solution [5]

• Missing values in numerical attributes were managed by replacing them with the mean value specific to each attribute. For nominal attributes, a mode imputation tech nique was utilized, where missing values were filled with the attribute's most frequent value. After completing this data preprocessing step [11], the data distribution underwent a transformation.

C. Machine Learning Models

• support vector machine (SVM)

Attributes	Description	Type of test	Attribute Type	Attribute values
age	Age	-	numeric	years
bp	Blood Pressure	-	numeric	mm/Hg
sg	Specific Gravity	Urine	numeric	1.005, 1.010, 1.015,
al	Albumin	Urine	numeric	0, 1, 2, 3, 4, 5
su	Sugar	Urine	numeric	0, 1, 2, 3, 4, 5
rbc	Red Blood Cells	Urine	nominal	normal, abnormal
pc	Pus Cell	Urine	nominal	normal, abnormal
pcc	Pus Cell Clumps	Urine	nominal	present, notpresent
ba	Bacteria	Urine	nominal	present, notpresent
bgr	Blood Glucose Random	Blood	numeric	mgs/dl
bu	Blood Urea	Blood	numeric	mgs/dl
sc	Serum Creatinine	Blood	numeric	mgs/dl
sod	Sodium	Blood	numeric	mEq/l
pot	potassium	Blood	numeric	mEq/l
hemo	Hemoglobin	Blood	numeric	gms
pcv	Packed Cell Volume	Blood	numeric	-
wbcc	White Blood Cell Count	Blood	numeric	cells/cumm
rbcc	Red Blood Cell Count	Blood	numeric	millions/cmm
htn	Hypertension	-	numeric	yes, no
dm	Diabetes Mellitus	-	numeric	yes, no
cad	Coronary Artery Disease	-	nominal	yes, no
appet	Appetite	-	nominal	good, poor
pe	pedal Edema	-	nominal	yes, no
ane	Anemia	-	nominal	yes, no
class	Class	-	nominal	ckd, notckd

Fig. 2. attributes and their respective values

Support Vector Machine (SVM) is a well-regarded and su pervised ML technique known for its capability to find out intricate patterns present in complex and noisy datasets [8]. It is widely used for binary classification tasks and is rooted in statistical learning theory. SVM employs various kernel functions to map samples that are not linearly separable in their original feature space to a higher-dimensional space, where they can be more effectively classified.

• Logistic Regression

Logistic Regression (LR) is basically an algorithm and a widely used classifier model which can help with respect to binary classification tasks. It aims to derive a method that can predicts outcomes for a binary dependent element or variable based on one or more independent elements. A key component of logistic regression is the sigmoid function, which outputs values within the range of 0 to 1. By applying a threshold value, typically 0.5, the classifier assigns the output to either class 1 or 0. In this framework, an input sample is determined as belonging to class 1 if the output exceeds 0.5; if it does not, it is assigned to class 0..

• K Nearest Neighbour

The k-Nearest Neighbors (KNN) algorithm is utilized to pre dict chronic kidney disease by assessing the medical attributes of patients. It operates by identifying the k nearest patients to a new patient and categorizing the disease status of the new patient based on the majority class among those k neighbors. In this context, it evaluates attributes such as blood pressure, serum creatinine levels, age, and more. KNN is a straightfor ward yet effective algorithm for disease prediction, leveraging the similarity of patient profiles to make predictions. However, the selection of the ”k” value and the distance metric used are critical factors influencing its accuracy.

D. specifics of Glomerular flitration rate(GFR)

The glomerular filtration rate (GFR) is a critical aspect which affects kidney function, measuring the rate at which blood is filtered per minute. It is estimated with high precision through comprehensive blood tests that assess

serum creatinine and serum cystatin C levels, along with factors like weight and age, contributing to a thorough evaluation of renal health [12]. This multi-parameter approach allows for a nuanced and accurate determination of GFR, providing valuable insights into kidney function.

Stage of Chronic Kidney Disease	Description	e-GFR Level
One	Kidney function remains normal but urine findings suggest kidney disease	90 ml/min or more
Two	Slightly reduced kidney function with urine findings suggesting kidney disease	60 to 89 ml/min
Three	Moderately reduced kidney function	30 to 59 ml/min
Four	Severely reduced kidney function	15 to 29 ml/min
Five	Very severe or end-stage kidney failure	Less than 15 ml/min or on dialysis

Fig. 3. stage classification of CKD

GFR plays a crucial role in predicting Chronic Kidney Disease (CKD) severity, as it assesses how effectively the kidneys filter waste products from the blood. A lower GFR indicates reduced kidney function, a hallmark of CKD. For instance, a GFR of 50 ml/min indicates the kidneys are filtering 50 milliliters of blood per minute. A GFR of 90 or above is considered normal, indicating healthy kidney function. As GFR decreases, it indicates declining kidney function, with a GFR below 15 indicating end-stage CKD [9].

Healthcare providers use GFR in conjunction with other clinical data to diagnose CKD, monitor its progression, and determine appropriate treatment plans. Regular GFR mea surements help in early detection of CKD, enabling timely interventions to slow disease progression and prevent com plications. In essence, GFR is a vital tool for assessing and predicting CKD by quantifying kidney function [13].

E. Remedy classification for stages

• Previous research has typically focused on developing an accurate classification model to predict the presence of Chronic Kidney Disease (CKD) in individuals [4], using various algorithms and datasets containing relevant attributes [5]. In contrast, this study extends beyond pre diction methods. Once CKD presence is predicted, it also recommends appropriate treatments or remedies based on the patient's current CKD stage, thereby addressing additional health needs.

III. DICUSSION

• The most important objective of this study was to identify crucial clinical test attributes for efficient computer-aided screening of CKD, while aiming to reduce diagnostic costs. Our framework's results demonstrate that machine learning (ML) models perform well in classifying CKD and non CKD cases, using a significantly fewer number of elements or attributes. The unnecessary use of a large number of test attributes can have significant financial im plications, which can hinder routine CKD screening. We evaluated three ML models using clinical test attributes to identify the most suitable model or classifier for accurate CKD diagnosis, by using essential attributes from either single clinical pathology, such as urine or blood, or both.

• Despite using a limited number of attributes, the ML models consistently achieved near-perfect accuracy. Among the three classifiers tested, logistic regression was identified as the most effective. This investigation into

ML techniques using reduced attributes emphasizes the potential for cost-effective CKD diagnosis [4]. Individu als who are initially screened for CKD using affordable urine pathology, along with commonly available tests like blood pressure, hypertension, and age, can be recom mended for further comprehensive CKD assessment.

• This early referral strategy, initiated by urine testing, pro motes subsequent evaluations integrating blood pathology and other pertinent attributes. This process facilitates automated ML-based diagnosis for the effective manage ment and treatment of CKD patients [4].

• In modern medical practice, transparency and the ability to interpret classifier decisions are crucial. Healthcare professionals need a clear understanding of the decision making process to establish trust in automated diag nostic systems. While traditional performance metrics like accuracy, sensitivity, and specificity assess model performance, they do not explain the roles and influences of specific attributes in decision-making. To enhance trust and confidence in ML-based CKD testing systems, explainable AI-based ML algorithms are integrated into it. These algorithms offer explanations and justifications for decisions, addressing the need for transparency and interpretability in the decision-making process [3].

IV. RESULT

The proposed diagnostic methodology for Chronic Kidney Disease (CKD) demonstrates feasibility in terms of both data imputation and sample diagnosis. After employing unsuper vised KNN imputation to address missing values in the dataset, the integrated model achieved a satisfactory level of accu racy. There is optimism that implementing this methodology in practical CKD diagnosis could yield favorable outcomes. Furthermore, this approach may have potential applications in the clinical data of other diseases within the domain of medical diagnosis [14].

Even then , it is quite important to acknowledge some limitations encountered during the overall model establishment process. The data samples which are available are relatively small, coming upto only 400 samples [5], which may con strain the model's generalization performance. Additionally, the dataset's binary categorization (CKD and not CKD) limits the model's ability to diagnose the severity of CKD.

To enhance the model's generalization performance and expand its diagnostic capabilities, future efforts will focus on collecting a larger and more diverse dataset [15]. This will involve gathering a wider range of complex and representative data to train the model more effectively. The current dataset, while sufficient for initial testing, may not fully capture the variability and intricacies found in real-world scenarios. By incorporating a broader spectrum of data, we aim to capture subtle patterns and nuances crucial for accurate diagnosis.

The initial goal is not only to upgrade generalization feature but also to help the model to accurately detect disease severity across different populations and conditions, thus addressing potential biases and ensuring the model's applicability to a wide range of cases. As the size and quality of the dataset increase, we anticipate the model will evolve and become more refined over time.

Moreover, the expanded dataset will support more robust validation processes, enhancing confidence in the model's predictions. This iterative process of data collection, model training, and validation will be critical in advancing the model's capabilities. Ultimately, the goal is to develop a highly reliable and an appropriate diagnostic asset suitable for widespread adoption in clinical settings, thereby significantly improving patient outcomes and advancing the field of medical diagnostics.

The algorithms used have provided us with following re sults.

A. logistic regression

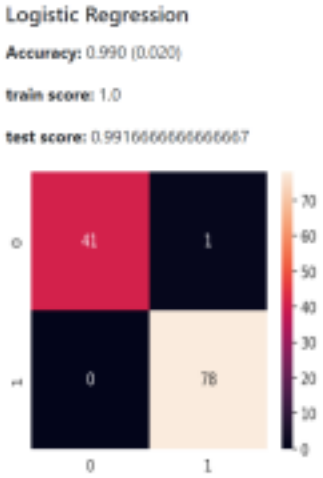

Fig. 4. logistic regression

A logistic regression model was trained and evaluated, achieving an accuracy of 0.990 (±0.020) on the test data. The model exhibited perfect accuracy on the training set (Train Score: 1.0) and a high accuracy of 0.992 on the test set. These findings demonstrate strong performance and effective generalization of the model .

B. K-nearest neighbours

The K-Nearest Neighbour (KNN) algorithm contributes to the CKD prediction model with an accuracy of 0.972 ± 0.029 and a test score of 0.975. The confusion matrix reveals 75 true positives, 42 true negatives, 3 false positives, and 0 false neg atives. These results demonstrate robust model performance [1], accurately identifying CKD patients with high accuracy and minimal errors, ensuring dependable diagnosis.

Fig. 5. k nearest neighbour

C. Support vector machine

The SVM algorithm makes a significant contribution to the CKD prediction model, achieving high accuracy (0.990 ± 0.020) and an excellent test score (0.975). The confusion matrix illustrates the model's robust performance with 78 true positives, 39 true negatives, 3 false positives, and 0 false negatives. These results ensure reliable and early CKD diagnosis with minimal misclassifications [8], highlighting the effectiveness of the SVM

approach.

Fig. 6. support vector machine

The result basically identifies if the patient has CKD or not and also specifies the stage with the probability of the patient being in that stage represented by either 1 or 0. this is shown in Fig 7.

Fig. 7. example-result of a patient with CKD stage 3

V. CONCLUSION

The current study has successfully identified a robust methodology for Chronic Kidney Disease (CKD) classifica tion and attribute selection, offering improved simplicity and cost-effectiveness. The approach involves a two-step process: firstly, training and selecting suitable classifiers, calculating feature importance, and deriving a reduced dataset based on pathological tests and measured feature importance. Secondly, these classifiers are trained with the reduced datasets and evaluated using test datasets.

The overall result or conclusion of this series of analysis shows that the identified important features align well with current clinical understanding. Notably, the logistic regression classifier emerges as particularly effective, achieving high classification accuracy, especially when using pathologically categorized attribute sets.

The proposed logistic regression classifier, in combination with the reduced test attributes, holds promise for reducing diagnosis costs and improving early treatment planning. More over, the study extends its impact by offering personalized treatment plans tailored to individuals diagnosed with CKD, classifying them based on the distinct stages indicated by the CKD predictor.

This comprehensive approach not only aids in cost reduction but also facilitates improved patient management through personalized treatment strategies. The inclusion of all relevant details in the diagnostic report significantly enhances reliabil ity and patient satisfaction. By providing a detailed breakdown of individualized treatment plans based on identified CKD stages, the study contributes to a more holistic and patient centric approach to healthcare.

In conclusion, the outlined methodology not only proves effective in CKD classification and attribute selection

but also holds promise for broader applications in healthcare, emphasizing personalized treatment and cost-effective diag nosis. The study's findings pave the way for enhanced patient care by introducing advanced diagnostic approaches that could revolutionize current practices in managing chronic diseases. This methodology has the potential to optimize treatment strategies, improve patient outcomes, and reduce healthcare costs, setting a new standard for personalized and efficient healthcare delivery.

REFERENCES

[1] A. M. Cueto-Manzano, L. Cortes-Sanabria, H. R. Mart ´ ´ınez-Ram´ırez, E. Rojas-Campos, B. Gomez-Navarro, and M. Castillero-Manzano, ´ "Prevalence of chronic kidney disease in an adult population," Archives of medical research, vol. 45, no. 6, pp. 507–513, 2014.[2] H. Polat, H. Danaei Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," Journal of medical systems, vol. 41, pp. 1–11, 2017.

[3] L. A. Inker, A. S. Levey, H. Tighiouart, T. Shafi, J. H. Eckfeldt, C. Johnson, A. Okparavero, W. S. Post, J. Coresh, and M. G. Shlipak, "Performance of glomerular filtration rate estimating equations in a community-based sample of blacks and whites: the multiethnic study of atherosclerosis," Nephrology Dialysis Transplantation, vol. 33, no. 3, pp. 417–425, 2018.

[4] S. Dhivya, D. Prabha et al., "A novel approach on chronic kidney disease prediction using machine learning," in 2022 International Conference on Advanced Computing Technologies and Applications (ICACTA). IEEE, 2022, pp. 1–6.

[5] R. Misir, M. Mitra, and R. K. Samanta, "A reduced set of features for chronic kidney disease prediction," Journal of pathology informatics, vol. 8, no. 1, p. 24, 2017.

[6] S. Hu, Y. Liang, L. Ma, and Y. He, "Msmote: Improving classification performance when training data is imbalanced," in 2009 second inter national workshop on computer science and engineering, vol. 2. IEEE, 2009, pp. 13–17.

[7] X. Liu, N. Li, L. Lv, Y. Fu, C. Cheng, C. Wang, Y. Ye, S. Li, and T. Lou, "Improving precision of glomerular filtration rate estimating model by ensemble learning," Journal of translational medicine, vol. 15, pp. 1–5, 2017.

[8] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, S. Alrashed, and S. O. Olatunji, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," Computers in biology and medicine, vol. 109, pp. 101–111, 2019.

[9] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration," Journal of biomedical informatics, vol. 53, pp. 220–228, 2015.

[10] M. A. Islam, S. Akter, M. S. Hossen, S. A. Keya, S. A. Tisha, and S. Hossain, "Risk factor prediction of chronic kidney disease based on machine learning algorithms," in 2020 3rd international conference on intelligent sustainable systems (ICISS). IEEE, 2020, pp. 952–957.

[11] M. Hosseinzadeh, J. Koohpayehzadeh, A. O. Bali, P. Asghari, A. Souri, A. Mazaherinezhad, M. Bohlouli, and R. Rawassizadeh, "A diagnostic prediction model for chronic kidney disease in internet of things plat form," Multimedia Tools and Applications, vol. 80, pp. 16 933–16 950, 2021.

[12] A. H. Anderson, W. Yang, C.-y. Hsu, M. M. Joffe, M. B. Leonard, D. Xie, J. Chen, T. Greene, B. G. Jaar, P. Kao et al., "Estimating gfr among participants in the chronic renal insufficiency cohort (cric) study," American journal of kidney diseases, vol. 60, no. 2, pp. 250–261, 2012.

[13] O. Niel, C. Boussard, and P. Bastard, "Artificial intelligence can predict gfr decline during the course of adpkd." American journal of kidney diseases: the official journal of the National Kidney Foundation, vol. 71, no. 6, pp. 911–912, 2018.

[14] M. J. Pencina, R. B. D'Agostino Sr, R. B. D'Agostino Jr, and R. S. Vasan, "Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond," Statistics in medicine, vol. 27, no. 2, pp. 157–172, 2008.

[15] A. B. Chapman, L. M. Guay-Woodford, J. J. Grantham, V. E. Torres, K. T. Bae, D. A. Baumgarten, P. J. Kenney, B. F. King Jr, J. F. Glockner, L. H. Wetzel et al., "Renal structure in early autosomal-dominant poly cystic kidney disease (adpkd): The consortium for radiologic imaging studies of polycystic kidney disease (crisp) cohort," Kidney international, vol. 64, no. 3, pp. 1035–1045, 2003.