International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 5 ǁ May 2023 ǁ PP. 411-418

Prediction of Diabetic Kidney Disease Using Machine Learning Algorithms

S.Dharun kumar
Department of Information Technology
Sri Ramakrishna Institute of Technology Coimbatore, India

N.Murali Krishna
Department of Information Technology
Sri Ramakrishna Institute of Technology Coimbatore, India

Mr.K. Sathyaseelan
Department of Information Technology
Sri Ramakrishna Institute of Technology Coimbatore, India

S.Saran
Department of Information Technology
Sri Ramakrishna Institute of Technology Coimbatore, India

Abstract-*DKD stands apart from other types of CKD to a number of distinguishing traits. Patients with DKD frequently have higher levels of anaemia than those with CKD without diabetes. Chronic kidney disease (CKD) is a global problem that claims a huge number of lives. It is also given for machine learning techniques to forecast chronic kidney disease using clinical data. Four master learning techniques are explored, including Decision Tree Classifier(DTC), Random Forest (RF), Linear Regression (LR), Gaussian Naive Bayes (LR). Thus, the system can be used to forecast CKD at an early level in a cost-effective way, which will be beneficial to developing and less developed nations.To improve the visualization of website smart framework is used.A smart framework using Flask is a web development framework that allows developers to create complex applications with ease.Flask is a micro web Framework written in Python.*

Date of Submission: 02-05-2023 Date of acceptance: 14-05-2023

I. INTRODUCTION

Basically, CKD and DKD are similar to each other that we will use the CKD information to forecast the DKD. There are five phases of CKD, and there are methods to delay or halt the progression of kidney failure. Chronic renal disease is a condition that causes a slow loss of kidney efficacy, which can lead to mortality, and it is a major public health issue globally, particularly in poor and middle-income nations. When a kidney has chronic kidney disease (CKD), it cannot properly filter blood and does not function as it should. targeted to identify diabetes patients with persistent kidney impairment using machine learning approaches. The two kidneys resemble beans and are each around the measure of a hand [1,2,3]. On either side of the spine, one is found fair underneath the thoracic cage. Removal of waste materials and extra fluid from the body through urine is the kidneys' primary job. Urine is created by a series of extremely complicated excretion and re-absorption processes. For the body chemical composition to remain stable, this procedure is required. The kidneys create hormone those that affect the operation of other organs while performing the vital management in salt, potassium, and acid level of the organism.

II. RELATED WORKS

Angier Allen , Zohora Iqbal , Abigail Green-Saxena , Myrna,Hurtado , Janaoffman , QingqingMao , Ritankar Das used XGBoost and Random forest algorithms used in this study are machine learning algorithms (MLAs) that predict the stages of diabetic kidney disease (DKD) within five years of a type 2 diabetes mellitus (T2DM) diagnosis[1]. Two MLAs were trained to forecast DKD severity stages, and their predictions were compared to the CDC risk score. The MLAs were validated using a hold-out test set and an external dataset obtained from several facilities. The study found that the MLAs performed better than the

CDC chance score in predicting any-stage DKD and more severe endpoints, with an AUROC of 0.75 and above 0.82, respectively.

J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z.[2] Ye presented Machine learning approaches for predicting the evolution of chronic renal disease are being compared and developed. used XGBoost , Random forest, Elastic net, Lasso Regression, Ridge regression, support vector machine , Neural networks, K nearest neighbor algorithms used in this study are machine learning algorithms (MLAs) that predict the severity of chronic kidney disease based on non-urinary clinical factors such as blood-derived tests and demographic characteristics. The models were assessed based on various metrics such as specificity, exactness, log-loss, and exactness.

Swathi Baby, P. & Panduranga, T. presented Vital, Statistical Analysis and Predicting Kidney Disease Using Machine Learning, J48 algorithms to analyse and predict diabetic kidney disease[3]. The renal disease data collection is studied using data mining classification algorithms.

El Houssainy Rady and Ayman Anwar presented Prediction various stages of renal disease via data mining algorithms used Probabilistic neural networks , multilayer perceptron , support vector machine , Radial Basis function to uncover
and extract hidden information from clinical and laboratory patient data, which can help doctors by maximising accuracy for illness severity stage detection.[4]

F. E. Murtagh, J. Addington-Hall, P. Edmonds, P. Donohoe, I. Carey, K. Jenkins, and I. J. Higginson presented Stage 5 chronic renal disease patients' symptoms in the month before death, treated without dialysis.[5]This paper presents presents findings from a longitudinal symptom survey conducted using the patient-completed To survey the predominance of side effects in patients with serious (Organize 5) persistent renal illness within the month some time recently passing, specialists utilized the Commemoration Indication Evaluation Scale-Short Shape (MSAS-SF).The comparison of symptom prevalence with that of advanced cancer patients was also calculated using the MSAS-SF.

W. Gunarathne, K. Perera, and K. Kahandawaarachchi presented Information analytics for unremitting kidney infection performance evaluation of machine learning classification approaches for illness categorization and forecasting (ckd)[6]. The algorithms used in the paragraph are Multiclass Decision Forest algorithm, which were used to build classification models to predict whether a patient has CKD or not. Multiclass Decision Forest algorithm is the algorithm that performed best for the smaller dataset with 14 attributes. The exactness of the model using this algorithm was reported to be 99.1%.

A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina presented utilising predictive analytics and machine learning, early diagnosis of chronic kidney disease [7]. A add up to of 4 machine learning-based classifiers have been examined.
The best performance results were sensitivity 0.9897, and specificity 1. The results of the experiment show that developments in machine learning, with the help of predictive analytics, constitute a viable environment for identifying intelligent solutions, which in turn demonstrate the predictive capability in the area of renal disease and beyond.

Chan L, Nadkarni GN, Fleming F, et al presented Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease[8].In this random forest model, clinical model are used. However, there is no mention of their individual accuracies. In arrange to foresee a composite result of eGFR decay of 5 ml/min per year, 40% maintained decrease, or kidney disappointment inside 5 a long time, a irregular timberland demonstrate was prepared, and its execution was compared with that of a clinical show and Kidney Malady.

Your kidneys clean your blood by eliminating surplus water and waste, which results in urine. The kidneys' ability to filter is indicated by their glomerular filtration rate (GFR)[9]. Nearly 90% of the estimated 37 million adult Americans who have chronic kidney disease (CKD) are ignorant of their illness. When caught early, kidney disease can be prevented in some significant ways.

It is difficult to obtain an accurate GFR level since measuring GFR (mGFR) is a difficult and drawn-out operation. It is therefore not practical for patients or professionals. For this reason, medical experts calculate GFR using a formula. CKD frequently exhibits no symptoms until the last stages of the condition. For the purpose of detecting CKD as soon as feasible, accurate estimates of GFR are crucial. A straightforward blood test that evaluates your creatinine levels is the typical method for estimating GFR.

The regular breakdown of muscle tissue and the digestion of food protein both produce creatinine, a squander item. In addition to CKD, additional factors that might affect creatinine levels include food, muscle mass (the weight of your muscles), malnutrition, and other chronic conditions.

III. EXISTING SYSTEM

Previous research used two distinct model types, random forests (RF), and gradient-boosted trees (XGB) to evaluate various tree-based strategies. RF fitted many decision trees to the data, which combined their forecasts, while XGB progressively fits trees that improve on prior errors.

IV. PROPOSED SYSTEM

This study aims to create a machine learning model that can identify diabetic kidney disease with high accuracy. Four classification methods are explored: DTC , RF, LR, and Gaussian Naive Bayes. The scikit-learn Python package was used to implement the article, and a dataset was collected to predict diabetic kidney disease. Precision and accuracy are used to evaluate the classifiers and algorithms.

CLASSIFICATION METHOD

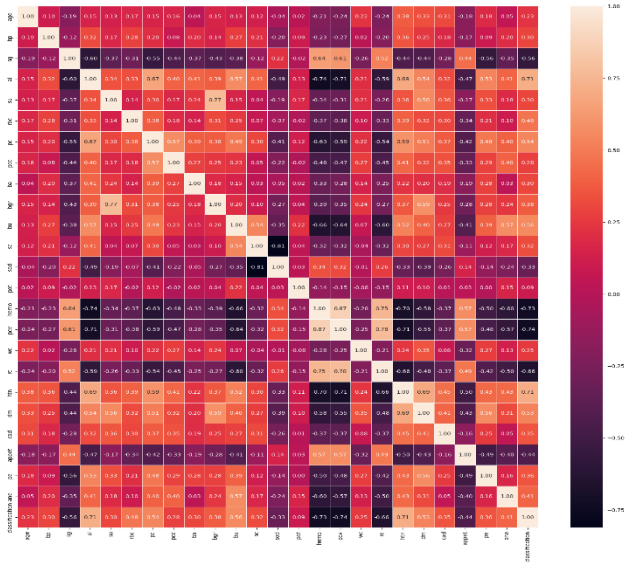
- Decision Tree Classifier
- Random Forest
- Gaussian Naive Bayes
- Linear Regression
- XG Boost

A. DATA COLLECTION

Data were collected over 2 months in India with 25 markers (eg red blood cell count, white blood cell count,etc.)inKagglewebsite(https://www.kaggle.com/datasets/mansoordaku/ckdisease). The goal is a "classification" that is either "ckd" or "notckd" - ckd=chronic kidney disease. There are 400 lines in total. The data has been cleaned in that it contains NaNs and the numeric functions need to be forced to float.

B. DATA PREPROCESSING

The algorithm such as Random Forest , Linear Regression ,Gaussian Naive Bayes , XGBoost , Decision Tree Classifier are used. The data is saved as a CSV file with 24 attributes and an output variable called "Class" with the value 'ckd' or 'notckd'. (Binary classification). This dataset also contains null values denoted by '- ', which we substituted with blank spaces in order for them to be rendered as null values in the pandas data frame. Because we have a reduced dataset of 400 records, instead of removing rows/columns with zero values, we replaced them with their corresponding column averages for numeric values and the string that has the greatest frequency in the respective columns for string values. We changed the text value comprising the columns to dummy/indicator columns after deleting the null values to make our data numeric and applying different machine learning models. (Binary columns). Methods such as the correlation matrix were used to apply another engineering trait. We taught the model properly by data is divided into test (30%) and training (70%). The training data is then partitioned three times into the training set and the validation set to compute the accuracy.



Correlation matrix

C.FEATURE EXTRACTION

Feature extraction could be a handle in machine learning that includes selecting and transforming the most relevant information from raw data to create a set of features that can be used as input for a model. This process is essential in many applications where the raw data is too complex or too large to be used directly by the model. By selecting and transforming the most important information, feature extraction can help improve the model's accuracy, reduce the amount of data required for training, and speed up the overall process. Include extraction procedures can shift depending on the sort of information being analyzed and the particular error at hand.

D.TRAINING MODEL

By dividing the data into test (30%) and training (70%) sets, we trained the model appropriately. After that, the training data are divided into three parts: the training set, the test set, and the reference set. the validation set to calculate the accuracy score. We also calculated the confusion matrix, precision, recall, f1-score, Feature importance and printed the parameter values of the models. ROC curves for models selected as the optimal subset of attributes for DKD prediction.

E.HYPERARAMETER TUNNING

Hyperparameter tuning is an important aspect of machine learning that involves finding the optimal values for the hyperparameters of model in order to improve its performance on the task at hand. In the case of predicting Diabetic Kidney Disease (DKD), there are several hyperparameters that can be tuned in order to improve the model's effectiveness.

One common approach to hyperparameter tuning is to use a grid search or a random search to explore the hyperparameter space. Grid search involves selecting a range of values for each hyperparameter and then testing each combination of hyperparameters to find the best combination that results in the highest performance. Random search involves randomly selecting values for each hyperparameter and testing a subset of the combinations to find the best performing set of hyperparameters.

1.DECISION TREE CLASSIFIER

A decision tree classifier is a popular algorithm used in machine learning for supervised learning tasksUp until a stopping requirement is met, the input data are divided recursively into subsets based on the values of one of the input features.. The divides are determined using the feature that maximizes the separation between the different classes in the resulting subsets, using a measure such as information gain .The resulting decision tree is a series of nodes and branches that represent a set of rules to classify new data based on the values of its features. Decision tree classifiers have a few focal points, counting their interpretability and ease of use. However, they can suffer from overfitting if the tree is too complex or if the training data is noisy. Various techniques such as regularization can used to address this issue.

2.RANDOM FOREST

A procedure for gathering learning called Arbitrary Woodland is utilized for both classification and relapse applications.This procedure is known for its tall precision and is especially valuable for dealing with huge datasets. Irregular Timberland was created by Leo Breiman and is based on the thought of decreasing fluctuation by combining the expectations of numerous choice trees. At preparing time, Irregular Timberland builds a expansive number of choice trees and produces a lesson that speaks to the cruel figure for relapse of all the person trees or the mode of the classes. The calculation at that point chooses a subset of the foremost appropriate trees and combines them to attain the most excellent result.

3.GAUSSIAN NAIVE BAYES

Gaussian Bayes may be a prevalent calculation utilized for classification errands in machine learning. It expect that each include is free of the others given the course name, making it computationally proficient and well-suited for expansive datasets.

To classify a new instance, it calculates the conditional probability of each feature given a particular class label using the Gaussian probability density function. It then combines these probabilities using Bayes' theorem to compute the posterior probability of each class given the observed feature values.

One advantage of Gaussian Naive Bayes is its simplicity, which makes it a popular baseline algorithm. However, its assumption of feature independence may not always hold true, which can lead to suboptimal performance.

Overall, Gaussian Naive Bayes is a useful algorithm for classification tasks involving continuous variables that follow a Gaussian distribution, but its limitations should be taken into account when using it in practice.

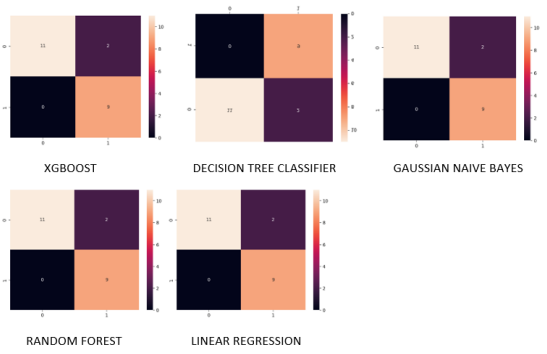
4.LINEAR REGRESSION

Direct relapse could be a factual strategy utilized to demonstrate the relationship between a subordinate variable and one or more autonomous factors. It includes finding a direct condition that can best clarify the relationship between factors. Basic direct relapse has one autonomous variable, whereas different straight relapse has more than one. Direct relapse is commonly utilized to analyze the relationship between factors and to create expectations. The quality of a direct relapse show can be assessed utilizing measurements such as coefficient of assurance (R-squared), cruel square blunder (MSE), and root cruel square mistake (RMSE).

5.XGBOOST

XGBoost (Extraordinary Slope Boosting) may be a well known machine learning calculation for classification and relapse issues. It is an outfit method that combines different frail forecast models to create a more grounded show. XGBoost employments a angle boosting system, where unused models are prepared based on the blunders made by the past models. The algorithm starts with a basic demonstrate, such as a choice tree, and after that iteratively includes more complex models to the outfit to move forward its performance.

One of the key highlights of XGBoost is its capacity to handle lost information, which could be a common issue in real-world datasets. It moreover bolsters diverse sorts of regularization methods, such as L1 and L2 regularization, to anticipate overfitting.XGBoost is known for its tall precision and quick preparing speed, making it a well known choice for numerous machine learning errands



Confusion matrix

F.PERFORMANCE METRICS

Performance metrics are used in machine learning to evaluate the effectiveness and accuracy of a model. Here are some commonly used performance metrics in machine learning:

Accuracy: It is the ratio of the number of correct predictions to the total number of predictions made by the model.

$Accuracy = TP + \ / TP + FP + FN + T$N

Precision: It is the ratio of the true positives (TP) to the sum of true positives and false positives (FP). It measures the ability of a model to correctly identify the positive class**.**

P$recision = TP / TP + T$N

Recall: It is the ratio of the true positives (TP) to the sum of true positives and false negatives (FN). It measures the ability of a model to identify all positive instances.

Re$call = TP / TP + F$N

F1 score: It is the harmonic mean of precision and recall, calculated as 2 * (precision * recall) / (precision + recall). It is a balance between precision and recall**.**

$F1\text{-}score = 2 *(precion * recall / \ precion + recall)$

Table 1
DECISION TREE CLASSIFIER

	Precision	Recall	F1-score	support
0	1.00	0.85	0.92	13
1	0.82	1.00	0.90	9
accuracy			0.91	22
macro avg	0.91	0.92	0.91	22
weighted avg	0.93	0.91	0.91	22

Table 2
RANDOM FOREST

	Precision	Recall	F1-score	support
0	1.00	0.85	0.92	13
1	0.82	1.00	0.90	9
accuracy			0.91	22
macro avg	0.91	0.92	0.91	22
weighted avg	0.93	0.91	0.91	22

Table 3
GAUSSIAN NAIVE BAYES

	Precision	Recall	F1-score	support
0	1.00	0.85	0.92	13
1	0.82	1.00	0.90	9
accuracy			0.91	22
macro avg	0.91	0.92	0.91	22
weighted avg	0.93	0.91	0.91	22

Table 4
LINEAR REGRESSION

	Precision	Recall	F1-score	support
0	1.00	0.85	0.92	13
1	0.82	1.00	0.90	9
accuracy			0.91	22
macro avg	0.91	0.92	0.91	22
weighted avg	0.93	0.91	0.91	22

Table 5
XGBoost

	Precision	Recall	F1-score	support
0	1.00	0.85	0.92	13
1	0.82	1.00	0.90	9
accuracy			0.91	22
macro avg	0.91	0.92	0.91	22
weighted avg	0.93	0.91	0.91	22

The report provides a summary of the model's accuracy, precision, recall, and F1 score for each class in the dataset. Accuracy measures the overall correctness of the model's predictions, while precision measures the proportion of true positives among all positive predictions, and recall measures the proportion of true positives identified by the model among all actual positives.

V. RESULT

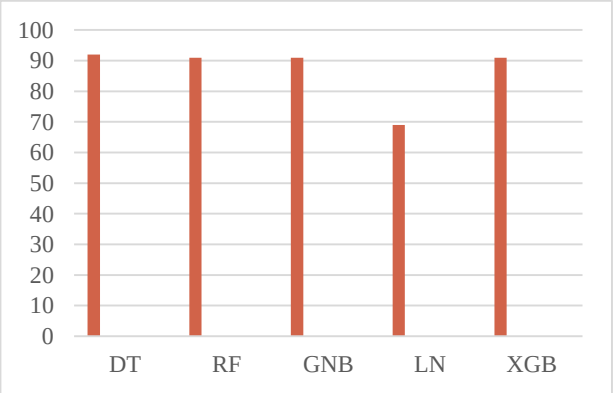
The entire outcomes of the experiment in terms of accuracy, precision, recall, and f1-score are presented above a table. For DT, RF,LR, GNB,XG the accuracy of these models is 92.82%,respectively. This table illustrates that Decision tree classifier provides highest level of accuracyand exceed the other approaches

A.Before Using Hyper parameter Accuracy Score

	Model	Tain Score	Test Score
1	Decision Tree Classifier	100.0	96.88
2	Random forest	100.0	100.0
3	Gaussian Naive Bayes	100.0	100.0
4	Linear Regression	94.7	88.77
5	XGBoost	100.0	100.0

B.After Using Hyper parameter Accuracy Score

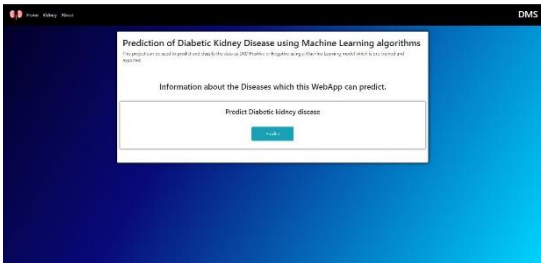
	Model	Train Score	Test Score
1	Decision Tree Classifier	92.82	86.36
2	Random forest	90.91	90.91
3	Gaussian Naive Bayes	90.77	90.91
4	Linear Regression	69.02	47.33
5	XGBoost	90.91	90.91

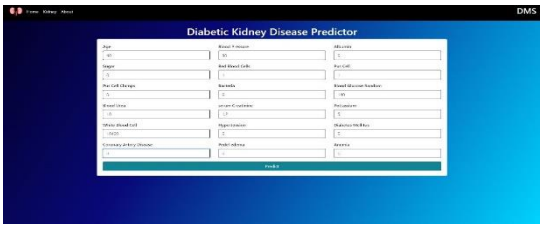


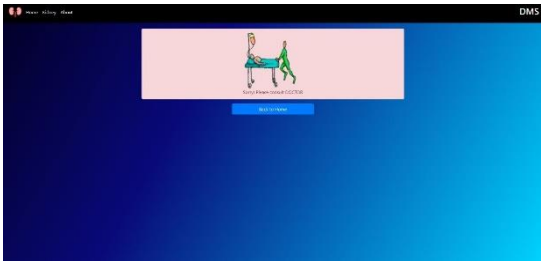
C.Web Development Using Flask

The Python web framework Flask was created to make it simple to create web apps rapidly and with the least amount of hassle. Because it doesn't come with many features by default but can be easily expanded with third-party packages, it is regarded as a micro-framework. For building web applications, managing HTTP requests, and providing static files, Flask offers an easy-to-use API. Additionally, it has a powerful templating engine and a built-in development server that make it simple to create dynamic content for your web pages. The adaptability of Flask is one of its main advantages. Everything from tiny single-page applications to expansive web services can be created using it. Additionally, it coordinating well with

other well-known Python libraries and frameworks, including Jinja2 for templating and SQL Alchemy for database integration. All things considered, Flask is a fantastic option for developers who want to easily and rapidly design web applications without having to deal with the complexity of a larger framework.







VI. CONCLUSION

The goal of this project is to track and evaluate the outcomes of applying various machine learning algorithms to the field of medicine in order to forecast diabetic kidney disease. This paper describes a forecast method for detecting DKD at an early stage. The models are trained and verified for the input parameters given, and the dataset contains input parameters collected from DKD patients. Decision tree, Random Forest, XGBoost, Linear Regression, Gaussian Naive Bayes algorithms are used to diagnose DKD. The precision with which models make forecasts determines their efficacy. The Gradient Boosting model, according to the study's results, predicts DKD more precisely than Decision Trees and Decision Tree Classifiers. This research's feature set selection and implementation duration, as well as its improvisation, can all be contrasted.

REFERENCES

[1]. Angier Allen , Zohora Iqbal , Abigail Green-Saxena , MyrnaHurtado , JanaHoffman , Qingqing Mao , Ritankar Das. "Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus", (2022).

[2]. J. Xiao, R. Ding, X. Xu, H. Guan, X. Feng, T. Sun, S. Zhu, and Z. Ye, "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of translational medicine, vol. 17, no. 1, p. 119, 2019.

[3]. Swathi Baby, P. &Panduranga, T. (2015). "Vital, Statistical Analysis and Predicting Kidney Disease Using Machine Learning Algorithms." International Journal of Engineering Research and Technology, 4(07), 206-210.

[4]. El Houssainy Rady and Ayman Anwar (2019). Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked. Volume 15, 2019, 100178

[5]. F. E. Murtagh, J. Addington-Hall, P. Edmonds, P. Donohoe, I. Carey, K. Jenkins, and I. J. Higginson, "Symptoms in the month before death for stage 5 chronic kidney disease patients managed without dialysis," Journal of pain and symptom management,

[6]. W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance evaluation on machine learning classification techniques for disease classification and forecasting through data analytics for chronic kidney disease (ckd)," in 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2017, pp. 291–296 .

[7]. A. J. Aljaaf, D. Al-Jumeily, H. M. Haglan, M. Alloghani, T. Baker, A. J. Hussain, and J. Mustafina, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," in 2018 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2018, pp. 1–9.

[8]. Chan L, Nadkarni GN, Fleming F, et al. Development and verification of a machine learning risk score for predicting the course of diabetic kidney disease utilising biomarkers and electronic patient data.. Diabetologia2021;64:1504–15.

[9]. "Estimated Glomerular Filtration Rate (eGFR)," . National Kidney foundation. Dec.2015