# Machine Learning and Deep Learning Application to Forecast Chronic Kidney Disease

Author; Owoeye Abiodun Oladayo
Artificial Intelligence and Data Science
Department of Physics,
Faculty of Science and Engineering,
Correspondence: a.o.owoeye-2021@hull.ac.uk

Supervising Lecturer:
Dr. Davis O Ojie Oseikhuemen
Department of Physics,
Faculty of Science and Engineering,
University of Hull, Cottingham Rd, Hull,
HU6 7RX, Kingston Upon Hull, United Kingdom.
Correspondence: o.o.ojie@hull.ac.uk

Director of DAIM: Dr. Kevin A Pimbblet
Department of Physics,
Faculty of Science and Engineering,
University of Hull, Cottingham Rd, Hull,
HU6 7RX, Kingston Upon Hull, United Kingdom.
Correspondence: k.pimbblet@hull.ac.uk

Domain Reviews: Dr. Eric Finlay
Consultant Pediatric Nephrologist,
Leeds Teaching Hospital NHS Trust,
email: eric.finlay@nhs.net
Dr. Anna Radford,
Consultant Paediatric Urologist and Surgeon,
Leeds Teaching Hospital, Leeds,
email: anna.radford@nhs.net

## Abstract

Machine learning and Deep learning implementation in healthcare have become a key part of technological advancement in the industry. Algorithms are widely used to improve decision-making processes such as early diagnoses and monitoring of chronic kidney disease (CKD). To diagnose CKD, A glomerular filtration rate (GFR) test is conducted. This test checks that the kidney is functioning properly. The result of the test is assessed with other human variables such as age, weight, and pre-existing conditions such as diabetes, hypertension, heart disease. The goal of this investigation is to build supervised learning and deep learning algorithms, integrated into a web application for a timely forecast of chronic kidney disease. The algorithms are logistic regression, random forest, kernel support vector machine, xgboost and four artificial neural network architectures with varying hyperparameters are evaluated and results are measured using statistical testing such as t-test, ANOVA, post hoc analysis (Bonferroni Correction), McNamar's test to determine the significance of their accuracies. The results of the hypothesis show that there is no statistically significant difference between the result of the models. Random forest appraised with the highest accuracy score 100%, ROC AUC score 100%, precision 100%, recall 100%, and F1-score 100%, outperforming the neural network with the deeper layer per accuracy 99%, ROC AUC Score 100%, precision 99%, F1-score 99%. During testing, the results also show that both classifiers appraised with the same proportion of errors. The Random Forest estimator is integrated and deployed as a web application. This study conclusively answers the question regarding timely diagnosis of CKD. The application can forecast CKD. Further studies are required to establish the device that can integrate this algorithm, accept blood samples for GFR test just like the blood sugar test device and make inference in relation with other predictors.

***Keywords:*** *chronic kidney disease (CKD), machine learning, deep learning, glomerular filtration rate (GFR), supervised learning, artificial neural network, hyperparameter, t-test, Bonferroni*

## Introduction

Artificial intelligence and its forms have become excellent techniques to utilize medical data in the most astounding ways; applying mathematics, computational statistics, and programming to build algorithms that can imitate patterns and aid disease detection in record time (Shailaja et al., 2018). The algorithms are trained with varying attributes from historical data. Pruijm et al., (2022) research highlights how Magnetic Resonance Imaging can be utilized to forecast diabetic renal disease; however, this detection technique is not widely applied in research and in clinical practice. There are many other contributors to disease detection which are not image-dependent but nominal or numerical in nature. The legitimacy of results has also been contended over time as Ahmad et al., (2018) argued that evaluation metrics alone are insufficient to explain the reason behind prediction to healthcare providers. The researchers argued that nuances are important in healthcare prediction. They highlighted a machine learning case study to predict pneumonia which predicted low-risk scores for patients with asthma but failed to consider that asthma patients are already on medication to lower symptom which inevitably rendered the data biased. Wong & Yip, (2018) study of cancer diagnosis explained that histology used to be the only procedure to identify cancerous tumors, machine learning algorithms can now analyse histology data. The experiment used samples of images of tumours, however, in a multi-label clustering classification of 82 classes, only 29 are recognised by WHO as a single category 1 tumour, others were subclasses (Capper et al., & Till et al., 2018). In summary, machine learning and deep learning are now a key part of healthcare, however, there are still questions about the legitimacy of the outcomes without clinical nuances.

## Chronic Kidney Disease

The kidney eliminates waste from the body (Albazi (2020). Chronic kidney disease is a collective term for a selection of conditions that impair the kidney's operation (Coresh, 2012). The most common causes are ureter damage, high blood pressure, diabetes, hypertension, and previous surgeries (Ene-Iordache, 2016). Soto (2012) study detailed a previous surgery's impact which documents a woman constantly obstructed left kidney due to recurrent hydronephrosis and ureter injury - the tube responsible for transporting urine Moinuddin (2015) which has been replaced 12 times with a ureteral stent following surgery 5 years prior. This condition in addition to diagnosis as hypertensive and diabetic degenerated into CKD. The common one in children is congenital ureteral stricture. Over the years, various machine learning approach has been proposed to aid early detection. For example, A representation of a Kidney is shown in Figure 1.

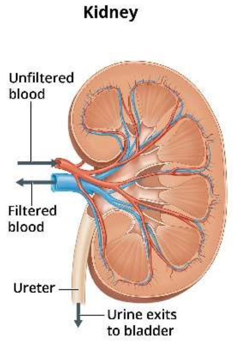


**Figure 1.** The Kidney, the bean-shaped clearing house, where blood is filtered and metabolic waste is disposed of (NIDDK, 2018)

## Background

Chen et al., (2016) applied a two fuzzy-logic approach to diagnose CKD. The researchers added noise to the data to reduce the memorization of training data, control overfitting and lower generalization (Brownlee, 2018). They created and trained a composite data with noise and also the original data. Comparing the results, sensitivity, and accuracy achieved an average prediction of 98% for both methods. When noise is added at random on the original data, and trained, they experienced improved accuracies of 99% on both methods.

Anusorn (2016) presented a machine learning study proposed Support Vector Machine, Decision tree, Logistic regression, and K-Nearest-Neighbor as estimators in addition, these was compared against to choose the most effective classifier. Experiments were assembled on Weka. Performance was measured by accuracy, sensitivity, and specificity. SVM demonstrated greater accuracy at 98.3% and sensitivity at 99%. Qin (2020) study implemented Support vector machines, logit regression, random forest, k-nearest neighbour, Naive Bayes, and feed-forward neural networks. The tree-based model demonstrated the best performance at 99.75%. They used perceptron to integrate random forest with logistic regression. as a neural network input. Results improved slightly to 99.83%. In another study, Amirgaliyev's (2018) paper proposed SVM as an effective method for class recognition because of its unique nature in finding the optimal hyperplane to accurately classify data points. Sequential Minimal Optimization was implemented to manage missing values and transform nominal features into binary. It achieved a sensitivity of 93.1%. CKD research in recent times has utilized the CKD data from the UCI repository, which has missing values. Aljaaf et al. (2018) addressed this by multiple imputations and subsequently built Multi-Layer Perceptron with four ML models, achieving an accuracy of 98.1%, 98.8% sensitivity, 100% specificity, 99.5% AUC. Vasquez-Morales et al. (2019) built a feed-forward network with a 95% validation accuracy.

Khan et al., (2020) experimented with Decision Tree, SVM, MLP, Naïve Bayes, LR, and Composite Hypercube on Iterated Random Projection (CHIRP). RMS Error, Mean Absolute Error, Root Relative Squared Error, Precision, recall, F-measure, Relative Squared Error, and Accuracy were the evaluation metrics. Equations 1-7 represents the formulae for the metrics and how they function.

$$RMSE = \sqrt{\overline{(f-o)^2}} \qquad \textbf{(1)}$$

f forecasts (anticipated values), o = known results. The mean is shown by the bar above the squared differences (similar to x).

$$\text{MAE} = \text{X}_{\text{experimental}} - \text{X}_{\text{true}} \qquad \textbf{(2)}$$

x-experimental is the measurement, x is the true value.

$$\text{RRSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \textbf{(3)}$$

number of observations, n
realized value, $y_i$
predicted value, $\hat{y}_i$
average of the realized values, $\bar{y}$

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction

$$Precision\ =\ TP/TP\ +\ FP \qquad \textbf{(4)}$$

$$Recall\ =\ TP/TP\ +\ FN \qquad \textbf{(5)}$$

Equations 4 and 5, True positive denotes the number of actual positive that predicted positive. False Positive implies the number of actual positives that predicted negative. False Negative represents the number of actual negative predicted positive.

Recall and precision are frequently at odds (Google Developer, 2022). Increasing precision decreases recall, and vice versa.

$$F - measure\ =\ TP/TP\ +\ ½(FP\ +\ FN) \qquad \textbf{(6)}$$

Equation 6. F-measure is dependent on precision and recall. It douses the tension between the two by providing a single score (Brownlee, 2020).

$$RSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (7)$$

Equation 7. illustrates the Relative Squared Error which shows the division of the MSE of the model by the MSE of the model which uses average as the predicted value.

$$Accuracy\ =\ TP\ +\ TN/TP\ +\ TN\ +\ FP\ +\ FN \quad \textbf{(8)}$$

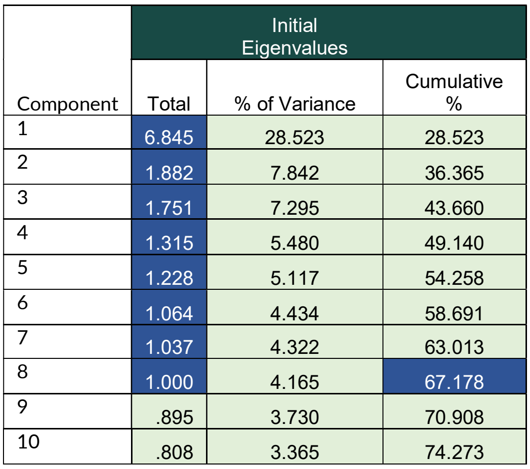
Equation 8. denotes the number of correct predictions divided by the total number of predictions.

In metrics as listed above CHIRP performed better overall with 99.75% accuracy. Nishat et. al., (2021) developed a robust diagnosis system for predicting CKD employing eight machine learning models. A comparative analysis was made with ROC AUC, precision-recall, sensitivity, and accuracy. Random forest performed best at 99.75%.

In this study, ML algorithms and ANN architectures were built to train 24 features. 10-fold cross-validation was employed across models for augmentation to improve how the model generalizes due to moderately small data. The evaluation metrics used are accuracy, ROC, AUC, and precision-recall curve. To lower model complexity, a factor analysis experiment was carried out using python and SPSS. The loaded weights, covariance matrices, of each attribute, and the rotated component matrices were also analyzed.

This revealed that dimensionality reduction operation to achieve the same level or optimal performance is eight (8). This was derived from eigenvalues ≥1 showing a variant of 67.178%. This research proposes DR at the point of prediction to address the difficulty in determining the features needed for deployment. Pairwise statistical tests were conducted for comparative analysis of 10-cross-validated results to establish statistical significance before the choice of model with the best performance. Random forest and ANN2 evaluated better overall.

Table 1.1 Total Variance Explained sums the variances of each individual primary component that makes up the overall variance explained.



| | Initial Eigenvalues | | |
|---|---|---|---|
| Component | Total | % of Variance | Cumulative % |
| 1 | 6.845 | 28.523 | 28.523 |
| 2 | 1.882 | 7.842 | 36.365 |
| 3 | 1.751 | 7.295 | 43.660 |
| 4 | 1.315 | 5.480 | 49.140 |
| 5 | 1.228 | 5.117 | 54.258 |
| 6 | 1.064 | 4.434 | 58.691 |
| 7 | 1.037 | 4.322 | 63.013 |
| 8 | 1.000 | 4.165 | 67.178 |
| 9 | .895 | 3.730 | 70.908 |
| 10 | .808 | 3.365 | 74.273 |

…

Lastly, chi-squared distribution was employed using the McNamar's test to compare ANN2 & RF.

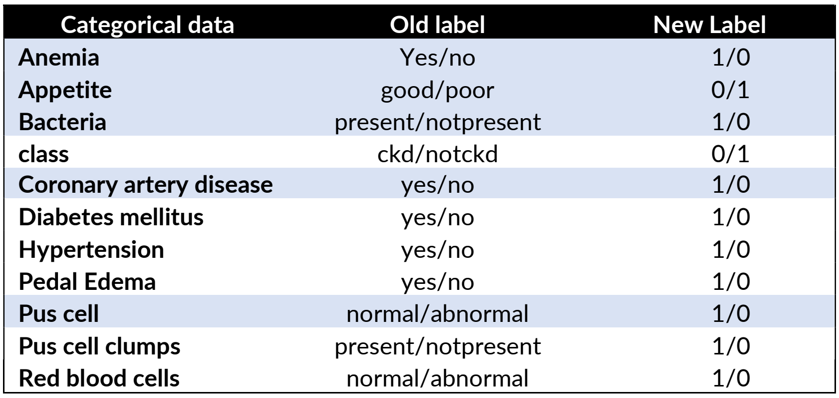
## Methods

### Data Collection and Pre-processing

The data was collected from "UCI Machine Learning Repository" (Soundara, 2015). Data consists of 14 Numeric features, 11 categorical features, and 400 observations. Attributes list can be found in Table 3.13. Missing values input were "?". This was replaced with the mode of each feature. Some attributes' mode was "?". This was replaced with the mode of the part of each series without "?". All categorical series were label encoded to binary and all elements were cast into their precise data types. A representation of how the categorical variables were encoded is shown in Table 2. Data was augmented using 10-fold cross validation.

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction

Table 2. 1 Encoding Categorical Variables



| Categorical data | Old label | New Label |
|---|---|---|
| **Anemia** | Yes/no | 1/0 |
| **Appetite** | good/poor | 0/1 |
| **Bacteria** | present/notpresent | 1/0 |
| **class** | ckd/notckd | 0/1 |
| **Coronary artery disease** | yes/no | 1/0 |
| **Diabetes mellitus** | yes/no | 1/0 |
| **Hypertension** | yes/no | 1/0 |
| **Pedal Edema** | yes/no | 1/0 |
| **Pus cell** | normal/abnormal | 1/0 |
| **Pus cell clumps** | present/notpresent | 1/0 |
| **Red blood cells** | normal/abnormal | 1/0 |

### Feature Selection

Employing correlation, data showed strong multicollinearity e.g., Hemoglobin showed a high positive relationship with 11 other attributes which suggests that some of the features can be excluded given that they will be doing the same job if implemented in a model.

### Dimensionality Reduction

Principal Component Analysis operation was done to extract the distribution that are highly significant and independent of one another to lower the bias in prediction, lessen the complexity of the proposed algorithm. The number of components was chosen based on the eigenvalues ≥1 .
The eigenvalues and eigenvectors were calculated from the decomposition of the covariance matrix Equation 9.

eigenvalue, eigenvectors =

np.linalg.eig( $cov(x) =$

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T) \qquad \textbf{(9)}$$

Splitting Data: Data was sliced into independent and dependent matrices of features and divided into 70%-30% training and test data.

Balancing Dataset: 62.47% of the data is “ckd”, 37.53% is “not ckd”. Data was balanced using smote to perform synthetic minority oversampling to reduce the bias of prediction (Chawla et. al., 2021).

Feature Scaling: Data was normalized using the standard scaler.

## Experiment and Results

### Model and Parameters

**Logistic Regression*:*** The odds are transformed using a logit function, which divides the odds of success by those of failure. Equation **(10)** and **(11)** mathematically represent this logistic function:

$$Logit(pi) \;=\; 1/(1+\;exp(-pi)) \qquad \textbf{(10)}$$

$$\ln\left(\frac{pi}{1-pi}\right) \;=\; Beta_0 +\; Beta_1 * X_1 +$$

$$\ldots\; +\; B\_k * K\_k \qquad \textbf{(11)}$$

logit(pi) is the label, X is the independent features. $Beta$ is the maximum likelihood estimation (MLE). To get the best fit of log odds, this technique iteratively examines various beta values. This training result average is 99.11% in a 10-fold cross-validation fitting and 99.16% validation. The model is a good fit.

**Kernel SVM:** A unique estimator to find the optimal hyperplane. The kernel is adjusted to manage non-linearity and higher dimensions of data (scikit-learn, 2022). The kernel utilized is the radial basis function “rbf” which was used to control the higher dimension. Equation (12) describes SVM:

$$\min_{w,b,\zeta}\frac{1}{2}w^Tw + C\sum_{i=1}^{n}\cdot\zeta_i$$

$$.$$

$$. \qquad \textbf{(12)}$$

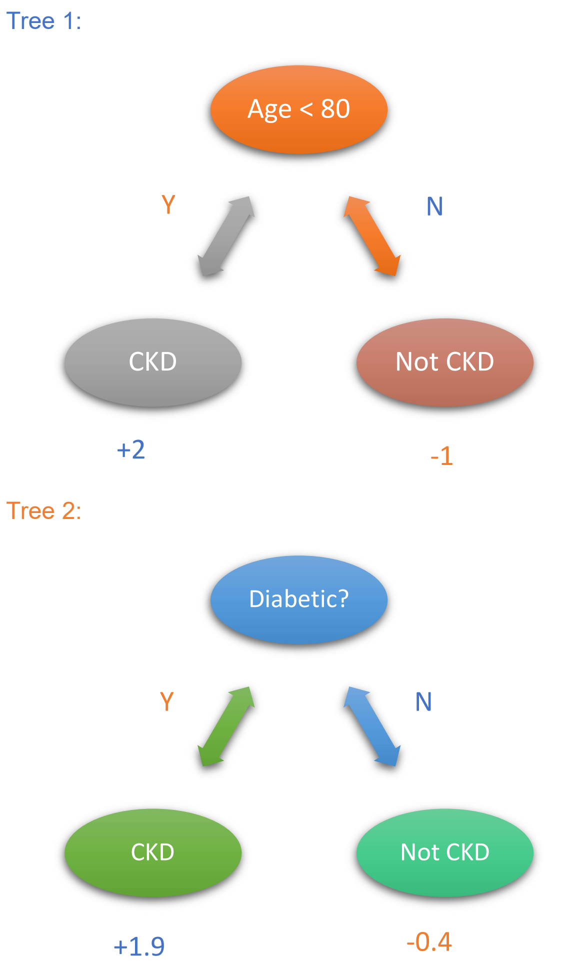
subject to $\quad y_i(w^T\phi(x_i)+b) \geq 1-\zeta_i,$

$$\zeta_i \geq 0, i = 1,\ldots,n$$

The result of training averaged 98.52% and tests at 98.3% which shows that the model is a good fit.

**XGBoost:** An optimized distributed algorithm that implements the gradient boosting framework. It classifies the training data into different leaves and assigns them the scores on the corresponding leaf. The operation takes every variable in a tree-like pattern, assigns scores and performs a linear algebraic calculation to classify the results as illustrated in Figure 2.

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction



f(CKD) = 2 + 1.9=3.9 f(Not CKD) = -1 – 0.4 = -1.4

**Figure 2:** Xgboost Process

Our model can be expressed in Equation (13):

$$\hat{y}_i = \sum_{k=1}^{K} \cdot f_k(x_i), f_k \in \qquad \textbf{(13)}$$

A function in the functional space $\mathcal{F}$, is called $f_k$ . $\mathcal{F}$ is the set of all possible trees. Quantity of trees $K$

The number of trees for this operation is 250. The accuracy after training was 97.65% and test accuracy 99.16%. The model is a good fit.

**Random Forest:** This estimator that averages the results of several decision tree classifiers fitted to different subsamples of the dataset to increase predicted accuracy and reduce overfitting (scikit-learn, 2022). The max feature preferred is 4 for the best split. In a classification exercise, the projected class will be determined by a majority vote (IBM Cloud Education, 2020).

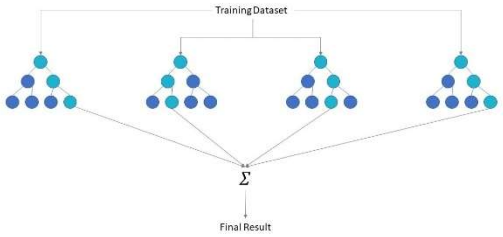


**Figure 3.** Random Forest combines outputs of multiple decision trees to arrive at a single outcome (IBM Cloud Education, 2020)

The estimator was trained and validated with a 100% score.

Table 3. 1 Supervised Learning Accuracies

| | Logistic Regression | Kernel SVM | XGBoost | Random Forest |
|---|---|---|---|---|
| | 1.0 | 1.0 | 1.0 | 1.0 |
| | 1.0 | 1.0 | 0.97 | 1.0 |
| | 1.0 | 0.97 | 1.0 | 1.0 |
| | 0.94 | 0.94 | 0.97 | 1.0 |
| | 0.97 | 0.94 | 1.0 | 1.0 |
| | 1.0 | 1.0 | 0.91 | 1.0 |
| | 1.0 | 1.0 | 0.97 | 1.0 |
| | 1.0 | 1.0 | 1.0 | 1.0 |
| | 1.0 | 1.0 | 0.97 | 1.0 |
| | 1.0 | 1.0 | 0.97 | 1.0 |
| **Average** | 0.991 | 0.985 | 0.976 | 1 |
| **AVG(STD)** | 0.006 | 0.008 | 0.009 | 0 |

The comparative supervised learning Model Accuracy Figure 4:

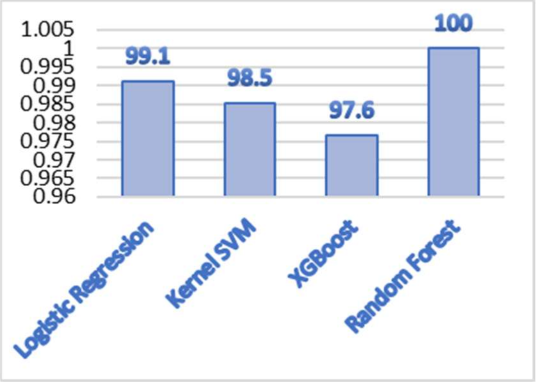


**Figure 4.** Supervised Model Accuracy Comparison

### Statistical Testing of Supervised Learning Accuracy

The test proposes to prove if accuracies are the same H1 or there are noticeable differences $H^{θ}$.

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction

### Hypothesis 1

Table 3. 2 ANOVA Result below describes the test for the overall difference between the groups.

| Source | ddof1 | ddof2 | F | p-unc | np2 | eps |
|---|---|---|---|---|---|---|
| Within | 3 | 27 | 2.4 | 0.08 | 0.21 | 0.48 |

**Result:** Test fails to reject null ($H^{\theta}$). The p-value > α (0.05), reveals that there is “no statistical significance between the accuracy”.

### Hypothesis 2

Equation (14) describes a six-pair combination derived using itertools.combinations as embedded with this equation 14.

$$C(n,r) = \binom{n}{r} = \frac{n!}{(r!(n-r)!)} \qquad \textbf{(14)}$$

Table 3. 3 Illustrates six family-wise t-test. One pair tests 0.01 which is less than α. The result is statistically significant and will require a post hoc test.

| Pairwise Combinations | P - values |
|---|---|
| Log Reg x KSVM | 0.5672 |
| Log Reg x XGB | 0.1846 |
| Log Reg x Random Forest | 0.1768 |
| KSVM x XGB | 0.4623 |
| KSVM x Random Forest | 0.0792 |
| XGB x Random Forest | 0.0133 |

### Post Hoc Analysis

Post Hoc is necessary when a test rejects the null.

**Bonferroni Test:** The Bonferroni correction involves generating a new alpha by dividing the default alpha by the number of groups. New_α= α/6 = 0.008. This test reveals how significant the difference in accuracy is if the group that rejects the Null is still less than the new alpha.
**Result:** Test (0.01 > 0.008) fail to reject null ($H^{\theta}$).

### Evaluation Metrics ML Models

#### Precision recall curve

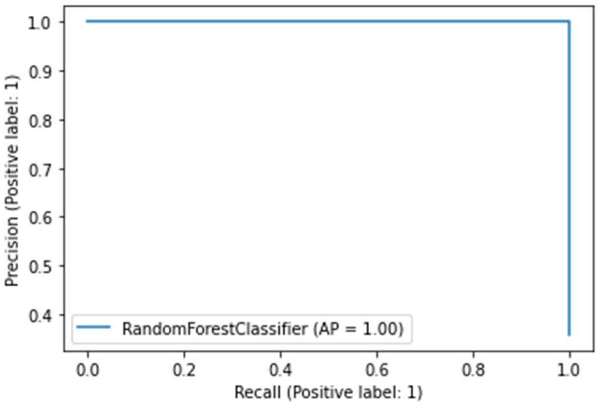


**Figure 5.** Illustrates the trade-off between precision and recall for various thresholds depicted by the precision recall curve. High precision corresponds to a low false positive rate, while high recall corresponds to a low false negative rate, and a high area under the curve denotes high recall and high accuracy(scikit-learn,2022).

Table 3. 4 Illustrates the accuracy, precision, and F1-score that defines the performance of the four ML models

| Model | Accuracy | Precision | F1-score |
|---|---|---|---|
| **LR** | 99.16 | 99 | 99 |
| **KSVM** | 98.30 | 98 | 98 |
| **XGB** | 99.16 | 99 | 99 |
| **RF** | 100.0 | 100 | 100 |

Table 3. 5 ROC AUC Score

| Model | ROC AUC Score |
|---|---|
| **LR** | 99.98 |
| **KSVM** | 99.84 |
| **XGB** | 98.53 |
| **RF** | 100 |

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction

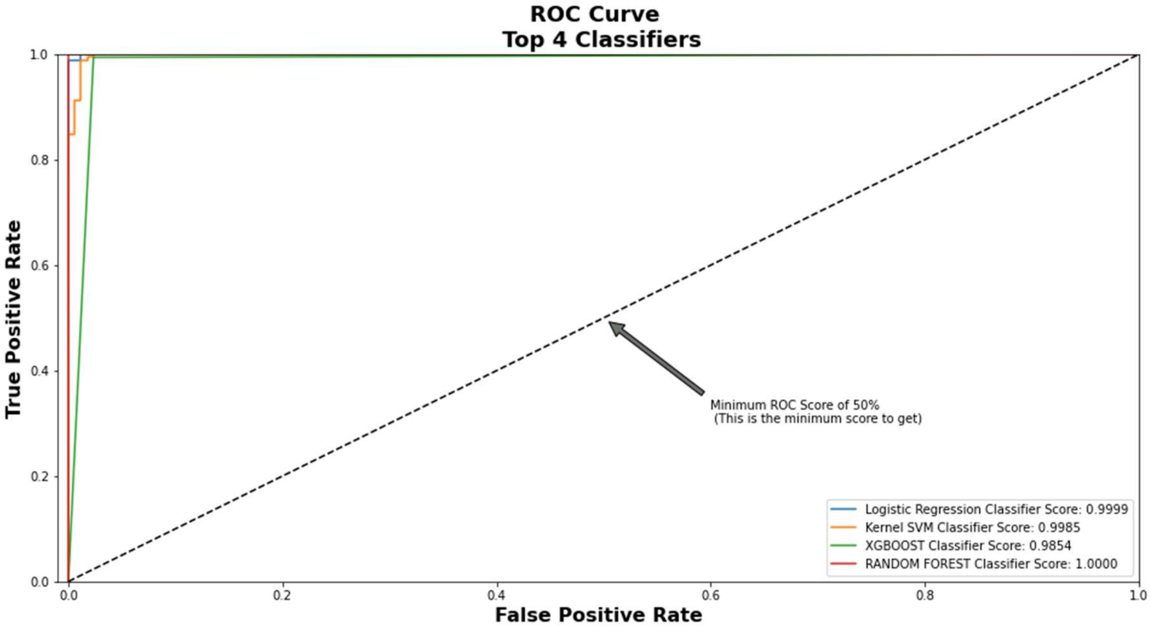


**Figure 6.** The graph above illustrates the ROC curve. Random Forest displayed the highest area under the curve.

### Neural Networks

The neural networks accuracy from the 10-fold cross validation fitted on the sequential model as consistent with the ML models. Scikit-learn Keras-wrapper, Keras Classifier was employed to extend permission with sklearn workflow. Data was passed into the networks in 32 and 64 batch sizes.

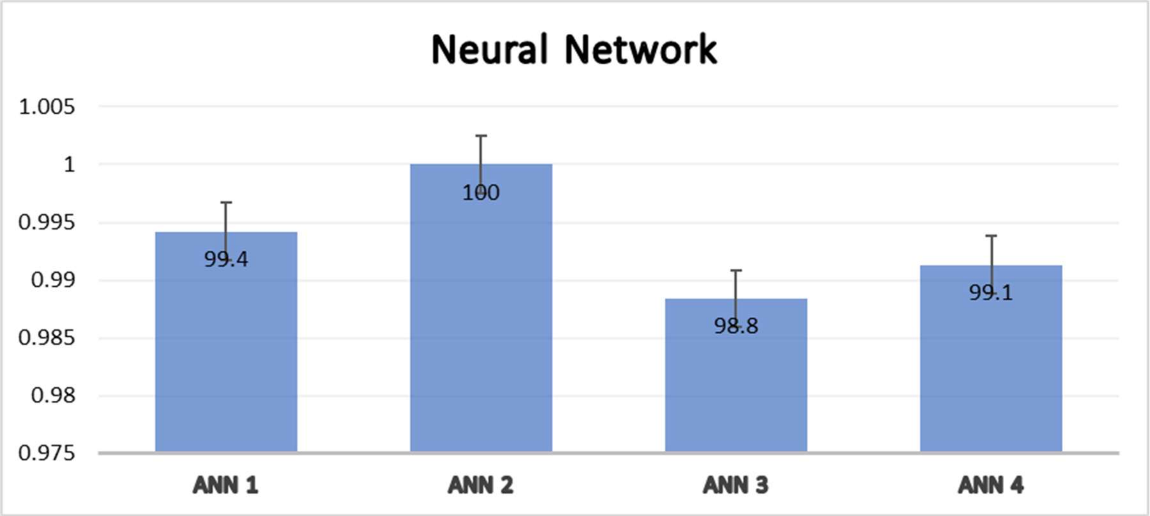


**Figure 7.** Neural networks accuracy

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction

### Artificial Neural Network Architectures

Table 3. 6 Describes the neural network model architectures and results

| Parameters | ANN_1 | ANN_2 | ANN_3 | ANN_4 |
|---|---|---|---|---|
| Activation function | Relu, Relu, sigmoid | Relu, Relu, Relu, tanh, tanh, sigmoid | tanh, tanh, sigmoid | Relu, sigmoid |
| Dense Units | 12, 8, 1 | 12, 8, 8, 8, 1 | 12, 8, 1 | 12, 1 |
| Epoch | 150 | 100 | 100 | 80 |
| Batch Size | 32 | 32 | 64 | 32 |
| Optimizer | Adam | RMSProp | SGD | Adam |
| Loss function | Binary_Crossentropy | Binary_Crossentropy | Binary_ Crossentropy | Binary_Crossentropy |
| | | | | |
| | ANN_1 | ANN_2 | ANN_3 | ANN_4 |
| | | | | |
| 0 | 1.0 | 1.0 | 1.0 | 1.0 |
| 1 | 0.97 | 1.0 | 0.94 | 0.94 |
| 2 | 1.0 | 1.0 | 1.0 | 1.0 |
| 3 | 1.0 | 1.0 | 0.97 | 1.0 |
| 4 | 1.0 | 1.0 | 1.0 | 1.0 |
| 5 | 0.97 | 1.0 | 1.0 | 0.97 |
| 6 | 1.0 | 1.0 | 0.97 | 1.0 |
| 7 | 1.0 | 1.0 | 1.0 | 1.0 |
| 8 | 1.0 | 1.0 | 1.0 | 1.0 |
| 9 | 1.0 | 1.0 | 1.0 | 1.0 |
| Average | **0.994** | **1.0** | **0.988** | **0.991** |
| AVG(STDEV) | 0.004 | 0.000 | 0.006 | 0.006 |

### Statistical Testing Neural Network Accuracy

Hypothesis 3:

**Result:** P-value > 0.05. There is "no statistical significance between the accuracy".

Table 3. 7 ANOVA

| Source | ddof1 | ddof2 | F | p-unc | np2 | eps |
|---|---|---|---|---|---|---|
| Within | 3 | 27 | 1.79 | 0.17 | 0.16 | 0.67 |

Hypothesis 4:

Table 3.8 illustrates six family-wise t-tests for the ANN results.

Table 3. 8 Test fails to reject null($H^{\theta}$)

| Pairwise Combinations | P - values |
|---|---|
| ANN 1 x ANN 2 | 0.1510 |
| ANN 1 x ANN 3 | 0.4465 |
| ANN 1 x ANN 4 | 0.6980 |
| ANN 2 x ANN 3 | 0.0852 |
| ANN 2 x ANN 4 | 0.1749 |
| ANN 3 x ANN 4 | 0.7431 |

### Evaluation Metrics ANN Models

#### Precision recall curve

The ANN architecture two displayed the best performance. It has a deeper layer than all other networks, this allowed the network to generalize better with more explaining power. This is implemented and described in the Figure 8.

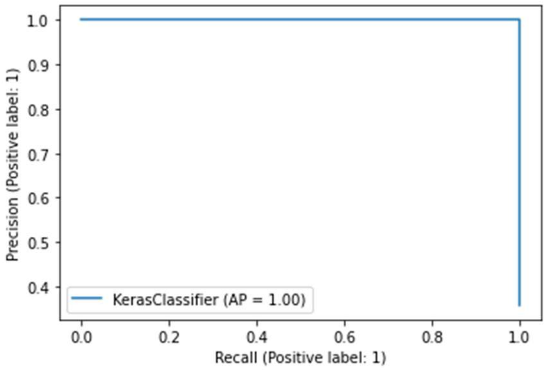


**Figure 8** is the precision-recall curve of neural network two with the highest accuracy.

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction

Table 3. 9 ANN Test Accuracy (%)

| Model | Accuracy | Precision | F1-score |
|---|---|---|---|
| **ANN 1** | 99 | 99 | 99 |
| **ANN 2** | 99 | 99 | 99 |
| **ANN 3** | 95 | 96 | 95 |
| **ANN 4** | 99 | 99 | 99 |

Table 3. 10 ANN ROC AUC Score

| **Model** | **ROC AUC Score (%)** |
|---|---|
| ANN 1 | 99.42 |
| ANN 2 | 100 |
| ANN 3 | 98.83 |
| ANN 4 | 99.12 |

### Comparing the ANN Model and Supervised Learning Model Results

#### Hypothesis 5

**Table 13.** ANOVA - ANN x SL Results below reveals that it fails to reject the null ($H^{\theta}$). There is no statistical significance between the accuracies. They are the same.

Table 3. 11 ANN X SL Results

| **Source** | **ddof1** | **ddof2** | **F** | **p-unc** | **np2** | **eps** |
|---|---|---|---|---|---|---|
| Within | 7 | 63 | 2.0141 | 0.0670 | 0.18 | 0.34 |

#### Hypothesis 6

ANN and SL accuracy were combined in 28 pairwise tests to determine the statistical significance. Xgboost x Random Forest and Xgboost x ANN 2 tests revealed a p-value of 0.01. This indicated that a post hoc test is required.

#### Bonferroni correction

new alpha = 0.05/28 = 0.0017. The p-values were checked against the 0.0017. This revealed that none is <= 0.0017.

Result: The test fails to reject the $H^{\theta}$, which suggests that the accuracy is the same.

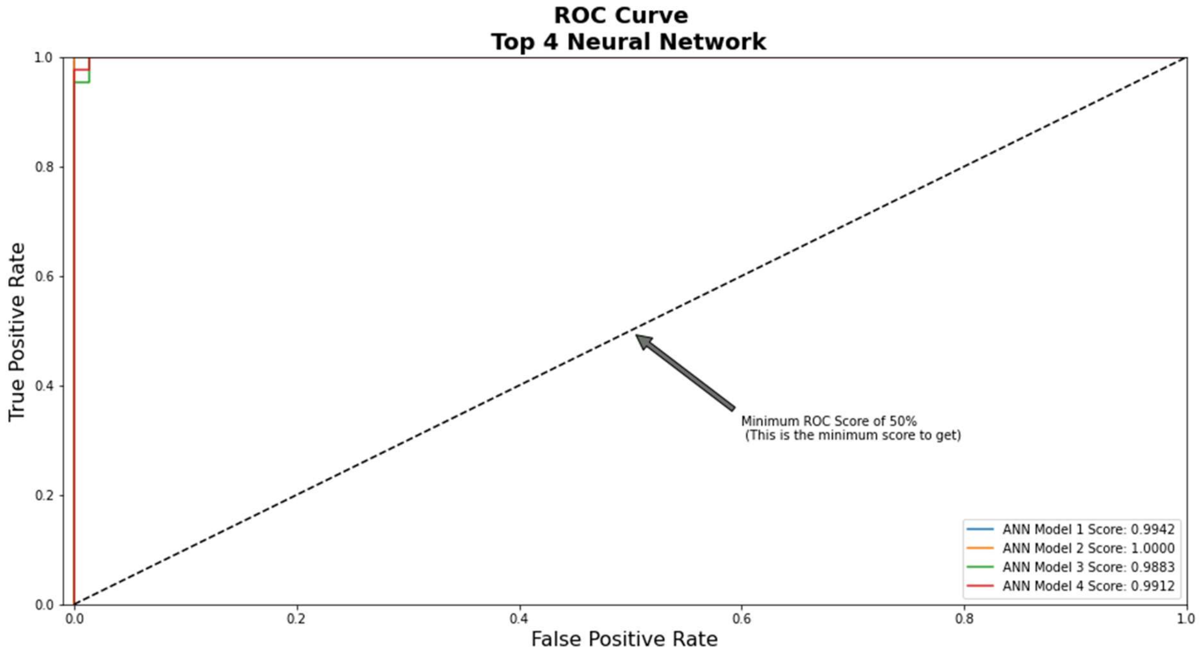


**Figure 9.** ANN 2 measured the highest area under the curve (AUC). The performance of the estimator at distinguishing between the +ve & -ve classes is improved by a larger AUC. (Huang & Ling, 2005)

### McNamar’s Test

A further test was conducted using the McNamar’s test which follows the chi-squared distribution. Here, we bring together RF and ANN 2 models to create a likelihood table which lists matrix of correct and incorrect decision of the test data. Equation 15 describes the McNamar’s test:

$$\chi^2 = (b - c)^2 / (b + c) \tag{15}$$

Table 3. 12 Likelihood table

| | Classifier 2 True | Classifier 2 Inaccurate |
|---|---|---|
| **Classifier 1 True** | Yes/Yes | Yes/No |
| **Classifier 1 inaccurate** | No/Yes | No/No |

**Result:** The test reveals that both consist of the same proportions of errors (fail to reject $H^{\theta}$).

To reaffirm the result of the Mcnamar’s test, a t-test was also conducted, the p-value = 0.318 reveals that the model performance is the same.

### Deployment

Determining the predictors for deployment was imperative. PCA data has been passed during model training to shed the dimension from 24 to 8 and to find the best fitting line but the actual predictors could not be determined because of the redistribution of loadings after DR. The loading matrix revealed 17 elements that demonstrated a positive correlation between PC1 and PC2 which holds the highest loadings.
Eight PCA data was passed into all algorithms but during testing, a shuffle of eight original independent variables X from the 17 was inputted to make a single prediction, 40%-50% erroneous predictions of ten (10) validation tests per algorithm were observed (80 tests overall).

The rotated component matrix (Varimax) using SPSS was analyzed, adjusted to exclude variants with less significance to improve determinant selection. The variables retained on the table were deployed but the selection failed tests for deployment.

Solution

The models’ input was reviewed and built to take all 24 original features and then a sklearn pipeline at the point of prediction was set up to perform three steps:

A. normalize raw data,
B. perform PCA from 24 to 8 and
C. fit the estimator (RF).

These steps achieved the desired result during testing, achieving 100% correct prediction after 10 validation tests per each algorithm (scikit-learn, 2022).

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction

Table 3. 13 Rotated Component Matrix

| Component | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Hemoglobin | -.772 | -.254 | -.101 | -.192 | -.229 | -.099 | -.012 |
| Anaemia | .763 | .062 | -.080 | .033 | .051 | -.033 | .132 |
| Packed cell volume | -.747 | -.269 | -.145 | -.207 | -.270 | -.107 | -.034 |
| Red blood cell count | -.618 | -.238 | -.073 | -.178 | -.200 | -.234 | -.080 |
| Hypertension | .522 | .046 | .318 | .204 | .330 | .368 | -.001 |
| Blood pressure | .513 | -.026 | .370 | .028 | -.167 | -.182 | -.054 |
| Pus cell clumps | .166 | .734 | .020 | -.056 | -.028 | .292 | -.083 |
| Pus cell | .253 | .717 | .101 | .071 | .189 | -.058 | .177 |
| Bacteria | .008 | .663 | .023 | .033 | .035 | .015 | .001 |
| Albumin | .209 | .639 | .237 | .144 | .377 | -.106 | .085 |
| Specific gravity | -.306 | -.407 | -.338 | -.138 | -.245 | .073 | .121 |
| Sugar | -.015 | .151 | .826 | .013 | .040 | .102 | .181 |
| Blood glucose random | .089 | .127 | .803 | .002 | .083 | .161 | -.010 |
| Diabetes mellitus | .327 | -.019 | .563 | .185 | .348 | .311 | -.031 |
| Serum creatinine | .199 | -.003 | .023 | .830 | .027 | .073 | .261 |
| Sodium | -.181 | -.107 | -.054 | -.830 | -.018 | -.079 | .211 |
| Pedal edema | .145 | .126 | -.009 | .153 | .784 | -.010 | .187 |
| Appetite | -.311 | -.105 | -.019 | -.067 | -.612 | -.101 | .052 |
| White blood cell count | .001 | .105 | .179 | -.140 | .501 | .032 | -.144 |
| Age | .126 | .001 | .276 | .040 | .108 | .623 | .024 |
| Coronary artery disease | .060 | .259 | .148 | .291 | .041 | .547 | -.002 |
| Red blood cells | .122 | .364 | .219 | .285 | .179 | -.471 | -.044 |
| Potassium | .066 | .012 | .105 | -.026 | -.068 | .008 | .900 |

### Web application

A Python flask application was built using Visual Studio code environment to render the Html, CSS codes and model. Heroku cloud was used to deploy the application in real-time. Random Forest was integrated to make predictions.
*Link:*
https://ckd-prediction-application.herokuapp.com/

## Discussion

This study presents a practical, end-user-friendly application to forecast chronic kidney disease. Nearly all studies utilized the same CKD data from the "UCI Repository". This research compared to others highlighted how data preprocessing was achieved. The cleaning process was largely not documented in most of the other studies except Amirgaliyev's (2018) and Aljaaf et al. (2018) work implemented sequential minimal optimization (SMO) and multiple imputations respectively. The SMO is better used on sparsely large datasets, and might have influenced lower sensitivity experienced.

In Qin (2020) study, the similarity and differences between models was not extensively quantified, however, some good results were observed. Statistical testing would offer more assurance to define distinctions between accuracies.

In subsequent research, this study will significantly benefit from Chen et al., (2016) approach which

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction

applied the fuzzy rule expert systems. This method adds noise to the data and due to constant change caused by the addition of noise, the network's memorization of training samples is reduced, a less generalisation error, more stable network. Traditional multivariate classifiers cannot match the advantages of fuzzy logic-based classifiers due to its robustness.

The Random Forest and ANN 2 algorithms demonstrated better performance with high sensitivity, and specificity to predict CKD with one common restraint, this study agrees with Ahmad et al., (2018) work that highlighted the importance of clinical nuances in addition to other model evaluations such as AUC, precision-recall. Also, in the researcher's exchange with domain experts, it reflects that they also would agree with this submission. This study deems medical practitioners' validation important. It should be the final evaluation leg to prove the legitimacy of results if ever in doubt.

**Research Extension**

This research can be further extended to impact real-world applications by:

- Adapting the experiment to be integrated into a device like the "blood glucose monitor" where blood samples can be taken and tested to give values of variables needed and then prediction made in addition with other features such as age, albumin, anemia etc. in a one-does-all approach.
- Adding a final leg of evaluation by sending results to digital urology consultants in real-time.

**Conclusion**

This paper presents machine learning and deep learning methods to predict chronic kidney disease utilizing Logistic regression, kernel support vector machine, xgboost, and random forest classifiers and four neural network architectures. Overall, the results show that Random Forest evaluated with the highest accuracy, precision, recall, F-measure, ROC AUC score with 100% in all metrics and comparative analysis using statistical methods to determine the significance of the differences between results. In comparison with alternative experiment that incorporated fuzzy method, which evaluated with 98% accuracy, it relatively performed better with higher accuracy using 10-fold cross-validation to augment data, however, the expert building method has more robustness in its application. This study also revealed that nuances from clinical personnel are notwithstanding important in healthcare decision support systems. The results of this study show that the deeper neural network in ANN 2 and the random forest classifier both successfully predicted the presence of CKD, however, the random forest classifier outperformed the highest-performing neural network in terms of average, standard deviations, precision, recall, ROC AUC score, and F1-score. A web application to predict chronic kidney disease was deployed on Heroku cloud. The application can predict CKD.

**References**

Albazi, Wefak Jbori., (2020). *Renal (urinary) system nephron functions*. Anatomy & Physiology: The Unity of Form and Function, 3(1), 880.

Anusorm Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach and Nitat Ninchawee (2016). *"Predictive analytics for chronic kidney disease using machine learning techniques," 2016 Management and Innovation Technology International Conference (MITicon)*, 2016, MIT-80-MIT-83, doi: 10.1109/MITICON.2016.8025242.

Andrew S Levey, Josef Coresh, (2012). *Chronic kidney disease,*The Lancet, Volume 379, Issue 9811,2012, Pages 165-180, ISSN 0140-6736, https://www.sciencedirect.com/science/article/pii/S0140673611601785. https://doi.org/10.1016/S0140-6736(11)60178-5.

Ahmed J. Aljaaf, Dhiya Al-Jumeily, Hussein M. Haglan, Mohamed Alloghani, Thar Baker, Abir J. Hussain, and Jamila Mustafina, (2018). *"Early prediction of chronic kidney disease using machine learning supported by predictive analytics,"* in PROC. IEEE Congr. Evol. Comput. (CEC), Jul. 2018, 1–9.

Ahmed J. Aljaaf, Dhiya Al-Jumeily, Hussein M. Haglan, Mohamed Alloghani, Thar Baker, Abir J. Hussain, and Jamila Mustafina, (2018). *"Early prediction of chronic kidney disease using machine learning supported by predictive analytics,"* in PROC. IEEE Congr. Evol. Comput. (CEC), Jul. 2018, 1–9.

Bilal Khan, Rashid Naseem, Fazal Muhammad, Ghulam Abbas and Sunghwan. Kim, (2020) *"An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy,"* in *IEEE Access*, vol. 8, 55012-55022, DOI: 10.1109/ACCESS.2020.2981689.

Bogdan Ene-Iordache, Norberto Perico,Boris Bikbov, Sergio Carminati, Andrea Remuzzi,Annalisa Perna, Nazmul Islam, Rodolfo Flores Bravo, Maurizio Gallieni,Igor Codreanu, Ariunaa Togtokh,Sanjib Kumar Sharma, Puru Koirala, Samyog Uprety, Ifeoma Ulasi, Giuseppe Remuzzi., (2016). Chronic kidney disease and cardiovascular risk in six regions of the world (ISN-KDDC): a cross-sectional study. ScienceDirect, 4(5), 307-319.

Capper, D., Jones, D., Sill, M. et al. (2018). DNA methylation-based classification of central nervous system tumours. Nature 555, 469–474 https://doi.org/10.1038/nature26000

Gabriel R. Vasquez-Morales, Sergio M. Martinez-Monterrubio, Pablo Moreno-Ger, and Juan A. Recio-Garcia, (2019) ''Explainable prediction of chronic renal disease in the Colombian population using neural networks and case-based reasoning,'' IEEE Access, vol. 7, 152900–152910, 2019.

Google Developers, (2022). *"Classification: Precision and Recall".* Available at:

https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall [Accessed 19 August 2022].

Google Developers, (2022). *“Classification: Accuracy”.* Available at: https://developers.google.com/machine-learning/crash-course/classification/accuracy [Accessed 19 August 2022].

IBM Cloud Education, (2022). *Random Forest*. Available at: https://www.ibm.com/cloud/learn/random-forest [Accessed 01 August 2022].

Jiongming Qin, Lin Chen, Yuhua. Liu, Chuanjun Liu, Changhao Feng and Bin Chen,(2020) *"A Machine Learning Methodology for Diagnosing Chronic Kidney Disease,"* in *IEEE Access*, vol. 8, 20991-21002, 2020, doi: 10.1109/ACCESS.2019.2963053.

Jason Brownlee, (2018). *“Train Neural Networks With Noise to Reduce Overfitting”.*Available at: https://machinelearningmastery.com/train-neural-networks-with-noise-to-reduce-overfitting/ [Accessed 20 August 2022].

Jason Brownlee, (2018). *“How to Use ROC Curves and Precision-Recall Curves for Classification in Python”.*Available at: https://machinelearningmastery.com/ROC-curves-and-precision-recall-curves-for-classification-in-python/ [Accessed 03 May 2022].

Jason Brownlee, (2018). *“How to Calculate McNemar’s Test to Compare Two Machine Learning Classifiers”.*Available at: https://machinelearningmastery.com/mcnemars-test-for-machine-learning/ [Accessed 15 June 2022].

Jason Brownlee, (2020). *“How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification”.* Available at: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/ [Accessed 15 June 2022].

Jin Huang and C. X. Ling, (2005) "Using AUC and accuracy in evaluating learning algorithms," in *IEEE Transactions on Knowledge and Data Engineering*, 17, 3, 299-310, March 2005, doi: 10.1109/TKDE.2005.50.

Jose, Soto Soto, Michael Phillips, Joseph Cernigliaro, and William Haley., (2012). Renal Autotransplantation for Iatrogenic High-Grade Ureteric Stricture. Hindawi - Case Reports in Urology, 2012(4), 3

K. Shailaja, B. Seetharamulu and M. A. Jabbar, (2018) *"Machine Learning in Healthcare: A Review,"* 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018,910-914, DOI: 10.1109/ICECA.2018.8474918.

Kim S, Lee W. (2017). *“Does McNemar’s test compare the sensitivities and specificities of the two diagnostic tests?”* Statistical Methods in Medical Research.;26(1):142-154. doi:10.1177/0962280214541852

Muhammad Fahad Zafar, (2017). “Convert-Arff-to-CSV”. Available at: https://github.com/mfahadzafar/Convert-Arff-to-CSV/blob/master/arffToCsv.py [Accessed 03 May 2022]

Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. (2018). *Interpretable Machine Learning in Healthcare*. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '18). Association for Computing Machinery, New York, NY, USA, 559–560. https://doi.org/10.1145/3233547.3233667

Muntasir Nishat M, Faisal F, Rahman Dip R, Nasrullah SM, Ahsan R, Shikder F, Ar-Raihan Asif MA-, Hoque MA. (2021). "*A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms"*. EAI Endorsed Trans Perv Health Tech [Internet].13[cited2022Jul.27];7(29):e1. https://publications.eai.eu/index.php/phat/article/view/1195

N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, (2021) “*SMOTE: synthetic minority over-sampling technique,”* Journal of artificial intelligence research, 321-357

National Institute of Diabetes and Digestive and Kidney Disease, (2018). *U.S. Department of Health and Human Services.* Available at: https://www.niddk.nih.gov/health-information/kidney-disease/kidneys-how-they-work [Accessed 06 August 2022].

Pruijm, Menno; Aslam, Ibtisam; Milani, Bastien; Brito, Wendy; Burnier, Michel; Selby, Nicholas.M. Selby; Vallée, Jean-Paul. *"Magnetic Resonance Imaging to Diagnose and Predict the Outcome of Diabetic Kidney Disease—Where Do We Stand?*" Kidney Dial. 2022, 2, 407–418. https://doi.org/10.3390/ kidneydial2030036

Scikit-learn, (2022). *Random Forest Classifier.* Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [Accessed 01 August 2022].

Scikit-learn, (2022). *Precision-Recall.* Available at: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html [Accessed 02 August 2022].

Simone Centellegher, (2020). *“How to compute PCA loadings and the loading matrix with scikit-learn”* Available at: https://scentellegher.github.io/machine-learning/2020/01/27/pca-loadings-sklearn.html [Accessed 15 July 2022].

Pypi, (2022). *“Scikit-Learn Wrapper for Keras”.* Available at: https://www.adriangb.com/scikeras/stable/migration.html [Accessed 20 July 2022].

Scikit-learn, (2022). S*klearn Pipeline.* Available at: https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html [Accessed 28 July 2022].
Scikit-learn, (2022). S*upport Vector Machine.* Available at: https://scikit- learn.org/stable/modules/svm.html [Accessed 29 July 2022].

Soundara Pandian, Jerlin Rubini, Eswaran Perumal, (2015). Chronic_Kidney_Disease Data Set. UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. Available at: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease [Accessed 3 May 2022].

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction

Yedilkhan Amirgaliyev, Shahriar Shamiluulu and Azamat Serek, (2018) *"Analysis of Chronic Kidney Disease Dataset by Applying Machine Learning Methods,"* IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, 1-4, doi: 10.1109/ICAICT.2018.8747140.

Zia, Moinuddin and Dhanda, Raman., (2015). Anatomy of the kidney and ureter. Nephrology, 16(6), 247-252.

Zewei Chen, Zhuoyong Zhang, Ruohua Zhu, Yuhong Xiang, and Peter B. Harrington, (2016) *"Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers"* Chemometrics Intel. Lab. Syst., vol. 153, 140–145, ISSN 0169-7439, https://doi.org/10.1016/j.chemolab.2016.03.004. (https://www.sciencedirect.com/science/article/pii/S0169743916300405)

Keywords: SL- Supervised learning, ML- Machine learning, TP – True Positive, FP – False Positive, RF - Random Forest, DR – Dimensionality Reduction