Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

A Novel Prediction Analytics Science of Kidney Disease based on Neural Network

Dr. K. Raja[1], Srinidhi Sathyamurthy[2*], Tanisha Garg[3], Pallavi Gupta[4]
[1]Professor of Head, Computer Science and Engineering, SRM IST,Ramapuram, Chennai, India
[2,3,4]SRM Institute of Science and Technology, Ramapuram, Chennai, India
*ss5772@srmist.edu.in

ABSTRACT

Chronic kidney disease (CKD) is a condition wherein people face a progressive deterioration of kidney function over a period. In medical fields, the prediction of this illness is perhaps one of the main critical things. Various automated ML algorithms can aid in the prediction of the patient's kidney condition and treatment. The importance of this disease's early determination approach requires attention, mainly in developing nations, where it is often identified late. So, the most sensible solution is to find a feasible solution and decreasing the disadvantages. The proposed system aims to introduce a novel method to use Feature selection by using Pearson correlation and Univariate selection. The implications of using clinical characteristics to detect victims with CKD using support vector machines procedure are investigated in this survey. For the dataset obtained out of a generic database, separate metrics had been used to compare algorithms. The results will then be compared with Neural networks. Through the proposed model, accurate prediction for CKD will be done, along with specificity and sensitivity. Proper comparison will also be done between existing and proposed algorithms.

Keywords
Chronic Kidney Disease (CKD), Neural network, Correlation Matrix, Univariate selection, Support Vector Classifier (SVC).

Introduction

CKD also referred to as Chronic Kidney Failure, is characterized by progressive organ dysfunction. Squanders and excess fluids from the bloodstream are channelled into the kidneys and excreted from the body. As chronic kidney disease advances, dangerously high volume of fluids, electrolytes, and wastes may build up in human blood. The chances of having Kidney diseases increase with age. According to a survey, African Americans, Indians, Americans tend to have a great risk for CKD. Within the early stages of chronic kidney illness, a person will have a few signs or side effects. Although in some cases early detection of CKD may not be possible as there won't be any symptoms. In such situations, a person can only know about the disease if they get their kidneys checked with blood and urine tests.

Chronic kidney disease can be very harmful if detected at a later stage. In [1], a new optimized and efficient algorithm for chronic kidney disease (CKD) titled Density- dependent Feature Selection (DFS) with Ant Colony Optimization Technique (D-ACO) was described. The preprocessing of data is done and then the wrapper method is used to generate the best set of features for further process. Then the combination of the DFS and ACO helps to diagnose the health of the patient using the medical data generated. This algorithm cannot be used in the future as the accuracy is not 100% and the algorithm used is too complex to understand and implement. Alzheimer's disease affects many people, an algorithm was devised using Neural networks in [2]. Various methods were used here, including the All-pairs technique. Here, all possible pairs of temporal data points were generated, which in turn meant that time was taken as a variable while plotting the points. But this cannot be used in real-time, as time is not fixed. In [3], a co-activation graph was used along with deep neural networks. Deep neural networks tend to have a

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

certain amount of complexity, that is, more than two layers. Graph analysis was further done to get the result, also giving the accuracy.

Optical Coherence Tomography is an innovative optical imaging model that renders high-resolution tissue pictures. It was used to calculate the number of glomeruli within a volume of kidney tissue in [4]. Various methods were used here, including the charged couple device camera and support vector machines. Likewise, the dataset is collected and filtered to get the required sets where a lot of time was consumed, and then the dataset after filtration was converted into a frame for comparison in [5]. Then a graph was plotted using the Random forest algorithm. But this approach cannot be adopted as the correctness is not high and chronic kidney disease is detected when the symptoms start being visible due to which the best treatment is delayed.

CKD may not end up apparent until the kidney functioning is entirely disabled. Treatment for this malady centers on reducing the movement of kidney harm, often by controlling the basic cause. If CKD advances, it can be lethal till dialysis is performed. Healthcare frameworks are beneath this and are facing greater challenges recently. It is constraining fundamental changes. So, a better and more effective system is needed, as a new wave of healthcare is being born. It's very challenging for the medical field to create a strong diagnostic method for diseases like CKD which would run faster and provide more stable results. The diagnostic technique used is much more complicated and thus there is a need to use computing methods like Neural Networks which can be utilized. For a disease like CKD, even partial detection can be much more helpful, and thus Neural Networks provide a reliable means for diagnosis.

The proposed model would be consisting of three main parts: Data collection and Preprocessing, Feature selection and Data cleaning, and Performance analysis. The proposed technique would intend to decrease or eradicate the obstacles which have already been notified. The Literature survey is discussed in Section II. In Section III, the Proposed design is described is explained. Results and discussion will be explained in Section IV. Section V will be the Conclusion which will be followed by Acknowledgements and References.

Literature Survey

There have been several techniques involved in predicting CKD. Machine learning is a popular method used for this as it is effective, and it is adaptable for a broad variety of conditions. Support Vector Classifier (SVC), Random Forest (RF), Neural Network (NN), Logistic Regression, as well as other Machine Learning techniques are primarily used in preventive analytics. A Neural network is an algorithm that is one of the most popular and widely used since the 1990s. Many people have worked on this and it continues to be used in various research and findings. In neural networks, we consider a network as a set of neurons and then perform various analyses on the same.

Neural Networks

A neural network is comprised of techniques that mimic the cognitive processes to find correlations in large datasets. In [6] Generalized Discrimination Value (GDV) was introduced which worked nicely with single-layerperceptrons but in multilayer perceptrons, a lot of error backpropagation was found. It had been used to calculate the magnitude of category distinction of named data point groups in high-dimensional areas, and the resulting curved graph represented the result. Though at the end after processing this method was able to handle data at multi-level

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

perceptrons there was a loss of performance and less accuracy so this method could not be used. Activation functions in neural networks are significant because they introduce non-linear properties to the networks so in [7] an activation function was proposed which was tested on four important datasets. Though the proposed method had shown an increase in classification and good performance of the functions as compared to the other existing functions still it cannot be used because the output of the functions are not verified in the hidden layers and also there was diverse behavior of the methods which were not investigated.
In [8], a deep learning method that used CNN was hypothesized to generate neural peak behavior from information supplied for information recorded with a generalized linear design to achieve a good relationship among the predicted and actual frequency of spiking in the output unit. But this method was not used as the method used has non-linearity with the higher-order kernels and so it was not fully efficient. In [9], a CNN was developed for differentiation of the intima-media complexity (IE) Region of Interest (ROI), with 450 samples used to train the network. However a completely autonomous technique for estimating IMT by a deep learning CNN methodology for initial CVD detection was explored, the data used was also very broad, and indeed the framework seemed to be a very time-consuming and lengthy method which was also quite cost-effective. A Deep Neural Network classifier coupled mostly with Discrete Wavelet Transform was used to classify a dataset of many brain MRIs into four categories in [10]. The proposed method required fewer hardware specifications and achieved good accuracy with large images, but the accuracy of the system could have been improved by the usage of CNN in the proposed model.

Chronic Kidney Disease

A recurrent time-domain estimate had been used to assess the acceptable method and additive distortion, and this design was also used to simulate a singular one-step regression variable for subjects with a huge database to generate a residual output in [11]. The restricted ARX model was being used to pursue biological trends uncovered using quadratic programming in another method. This algorithm was able to produce easy output coefficients and also report generation and data query was good but it cannot be used as it was less applicable for evaluation, does not work well with high dimensionality, and does not work well with a large dataset. The Chronic kidney disease dataset in [12] consisted of 24 attributes that were analyzed properly for the CKD detection. This method was able to provide an appropriate framework for modeling complex systems, explain the unexpected experimental data but it cannot be used as it takes an extremely long time to reach the stationary distribution and does not produce an exact answer plus can be computationally expensive.
In [13], the dataset consists of 400 samples where each set has 24 predictive variables/features where many values are missing and then the KNN algorithm is used to select the K total specimens with the smallest Euclidean distance in each subject including incomplete data to measure the likelihood of kidney disease. Though this approach was able to ease the burden on pathologists, decrease evaluation variability and the model outperforms in precision and agility but it cannot be used as it does not classify remote spots and it does not feature correlated with graft outcome. Early diagnosis and successful medication seem to be the only cure for Chronic Kidney Disease because it progresses progressively. As a result, a correlation feature classification subset reviewer with a greedy step-by-step search system is used in [14] to use the Support Vector Machine classification algorithm for timely identification. This method is not efficient to use as it is a very intricate method and is not cost-effective.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

Chronic kidney disease is a very vital disease having a high economic cost to health systems and is usually asymptotic until a later stage where health becomes very critical so to determine the prevalence of CKD around the world broken down by point, place, age, and gender, literature searches in 8 databases was made in [15] for review. The method proposed cannot be put into use as the strategy's deliverable was not evaluated and also the costs associated with it were not taken into consideration. In [16] a screening review was performed for the determination of rate prevalence and hazard determinants of CKD. The method was quite effective and produced great timeliness in the battle of CKD but cannot be used as it did not produce 100% accuracy.

Prediction Analysis

Prediction Analysis is the examination of information, regression methods, and data science strategies to calculate the likelihood of possible results based on statistics to make the best predictions about what will happen mostly in long term. In [17] the scientific formula of logistics, the description of the optimization problem, the finding of the regression model by gradient descent approach, and improvement of the sigmoid mechanism and it also established a vehicle evaluation prediction model for a consumer to accept/reject a certain car. Though this model provided a good accuracy, it cannot be used as it is a complex method and it is too interdependent which can cause a lot of errors. A KNN methodology depending upon class participation and feature weighting was presented in [18] for comparing two accuracies and calculating the weighted gap, which is used to predict the desired tags of specimens. Though classification accuracy was improved, performance reliant on category precision when dealing with sets of data with multi-dimensional features was a problem. By the usage of 311 classification datasets the evaluation of anticipating error rates, number of variables, etc., is done in [19] using Random forest where the comparison of variable selection methods for different datasets is done using random forest. This method identified the preferable methods based on the datasets, but it cannot be used as the model was not an idea for noisy datasets.
In [20], a model was proposed based on feature classification least square support vector machine for the prediction of short-term wind power. After the data analysis, the DBSCAN method was used for the uncertainty of data, and then the proposed prediction model was compared with unclassified examples. The proposed model had improved accuracy and feasibility, but it was a complex method and was also not cost-effective, so it was not used. In [21], implementation of the Naive Bayes classifier was described using a general tool kit and the applicability and computations with multiple domains of classification were monitored using a dataset. Though this model was applicable for real-life applications, it cannot be used as it required a good knowledge of Python which is not easy for many people, so it is not efficient.

Chronic Kidney Disease Using Neural Networks

CKD affects almost ten percent of the grown-up populace in the society and to date, there is no cure for CKD, early detection is the only method to cure the disease. So, in [22], three ML procedures: logistic regression, feed-forward NN, and deep learning were used for finding the diagnosis and performance of CKD where it was found that the feed-forward method was best in terms of performance. But this model could not be used as the model is not efficient for large datasets. In [23] the performance of Artificial Neural Networks (ANN) was investigated concerning the estimation of CKD using data preprocessing, data conversion, and ANN for the establishment of mapping many clinical factors to the patient's survival. The proposed model

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

showed good performance and accuracy but was not used as the hyperparameters in it were manually chosen due to an imbalanced dataset which automatically drawbacks the whole performance.

There is a huge need to differentiate Non-Chronic Disease (NCKD) and Chronic Disease (CKD) to discover the wellness state of a subject. The study of Artificial Neural Networks (ANN) used in [24] is used to achieve this. Four attributes – Creatinine, Urea, Sodium, and Potassium are used to determine if the sufferer has CKD. The dataset is collected from a general hospital. [25] endorsed using an integrated hybridized Deeply Convolutional Neural Network (AHDCNN) for efficient CKD analysis. The performance of a CNN-developed model is achieved by decreasing the function factor. A supervised tissue classifier is built using these high-level properties. This classifier differentiates between both types of tissue. The Internet of Medical Things was being used as a medium (IoMT). It ends with the use of predictive modeling, which offers a successful process for assessing actionable insights and demonstrating the estimates of the future. To replace the anomalies throughout the current method, KNN imputation had been used, which chooses multiple valid studies from the closest parameters to analyze the incomplete information for each unfinished study. Missing values are extremely significant in healthcare circumstances That's because many ideals are frequently overlooked.

In the current system, every categorical attribute was formatted to enable the computation in a computer. All of the generalized data were transformed into attributes. Feature extraction was used to train the system using independent parameters Besides, by using KNN and FNN, categorical variables were converted to quantitative forms, and the entire collected data was then standardized.

Proposed Design

The proposed system consists of 3 main parts, Data Collection and pre-processing, Feature Selection and Model Prediction, and finally the Performance Analysis. The data collected is based on different attributes like blood pressure, RBC count, WBC count, haemoglobin, etc. Once all the data is collected it is saved in a .csv file containing only numerical data. After this, the data is processed, and the missing values are replaced with estimated values using KNN Imputer. Feature Selection is performed on the resultant dataset. Two different feature sets are used, that is, Categorical Features and Numerical Features. Quantile and Power Transformer along with Robust and Standard Scalar is used to map the original values to a more uniform and normal distribution and making it Gaussian-like along with scaling features using statistics that are robust to outliers and dividing all the values with Standard Deviation and standardizing the features respectively. Correlation between variables is estimated using Pearson, Kendall, and Spearman Matrix. Feature Selection is used to extract the important features, which are inputs for the model prediction. The NN model is created, and the dataset is trained, and the accuracy is predicted. Once the prediction is done, a Performance Analysis is performed on the predicted result.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

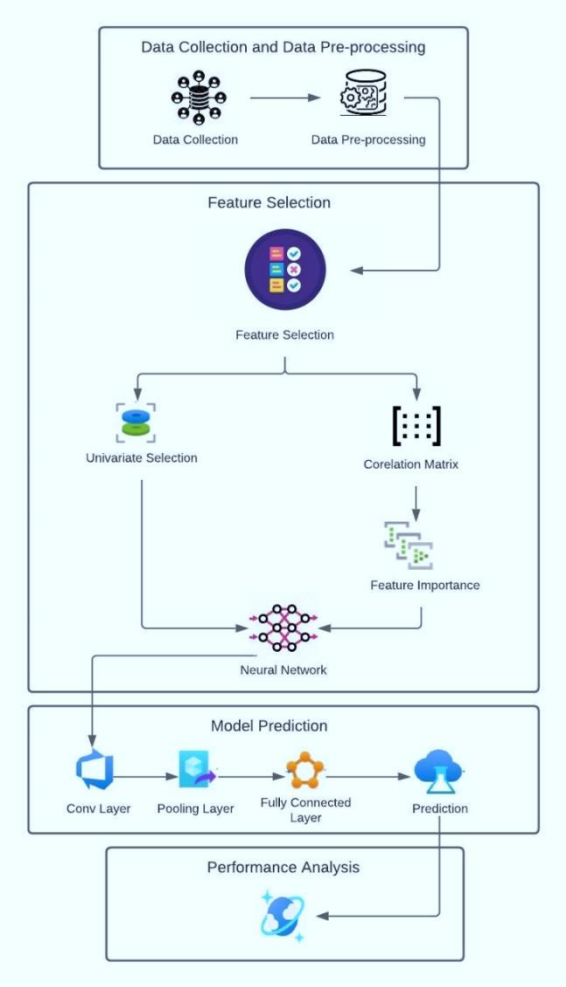


Figure 1. Architecture diagram of integrated model

As depicted in the Architecture diagram of the integrated model in Figure 1, the three modules used in the proposed algorithm are Data collection and pre-processing, Feature selection and model prediction, and Performance analysis.

The Pseudo- Code for the proposed algorithm is as follows:

Step 1: Start
Step2: A dataset (called “kidney_disease.csv”) is taken as the input for the proposed algorithm.
Step3: Extract the unique values for each column from the dataset.
Step4: Converting the data which is in String datatype to float datatype to maintain uniformity.
Step5: Quantile Transformer is used to transform all variables uniformly by mapping the original values to a uniform and normal distribution.
Step6: Power Transformer is used to make the data more Gaussian-like.
Step7: Robust Scalar and Standard scalar are used to scale features using statistics that are robust to outliers and dividing all the values with Standard Deviation and standardize the feature respectively.
Step8: Using KNN Imputer to find the missing values and replace them with the estimated value.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

Step9: Pearson, Spearman, and Kendall matrix is used to find the correlation between numerical variables(data)
Step10: K-Means algorithm is used to partition the dataset into pre-defined distinct clusters. By default, there are 8 K-Means Clusters.
Step11: Confusion Matrix is predicted after training the K-Means algorithm.
Step12: PCA (Principal Component Analysis) is used to find the pattern in the dataset and gives the variance for every component.
Step13: Training and predicting the result with PCA and T-SNE data.
Step14: Training and predicting the result with PCA + K-Means VS Target Variables.
Step15: Linear Discriminant Analysis is performed with Support Vector Machine by training the dataset first and predicting the result to find the accuracy score of the training set.
Step16: Different models are estimated and the correctness of training (during the training the percentage that the algorithm has learned) and testing (the percentage that the algorithm has predicted) sets are compared.
Step17: A model for Neural Network algorithm is created with epoch value as 1 that is, the number of times an algorithm will be trained. The dataset is trained and then predicted after which the accuracy of training and testing sets are compared.
Step18: A bar graph predicting the accuracy of training and testing is evaluated as output.

Data Collection And Pre-Processing

Data Collection is a process of gathering information that can be used to evaluate certain outcomes. Data can be collected from different sources like articles, journals, libraries (stored records), observations, surveys, interviews, etc. While building ML models, one has to give the most priority to data collection. For the dataset to be used in this research, we collected the following data of a person as shown in Figure 2.

age- age
bp - blood pressure
sg - specific gravity
al – albumin
su – sugar
rbc - red blood cells
pc - pus cell
pcc - pus cell clumps
ba – bacteria
bgr - blood glucose random
bu - blood urea
sc - serum creatinine
sod – sodium
pot – potassium
hemo – hemoglobin
pcv - packed cell volume
wc - white blood cell count
rc - red blood cell count
htn – hypertension
dm - diabetes mellitus
cad - coronary artery disease
appet – appetite
pe - pedal edema
ane – anemia
class – class

Figure 2. Data Collected

Data pre-processing is the methodology of organizing original information so that it can be used in an ML model. It is a technique for converting disruptive and large data into meaningful and clean records. In the normal community, the study is collected. Once the sample is collected, the data isn't clean and formatted. As a result, it may have inaccuracies, missing data, as well as

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

other skewed data. When a dataset is used for machine learning it is important to make sure that the data doesn't contain any missing value or invalid data, therefore Data pre-processing is done to achieve the same. No matter how well planned and well-designed our model is, it won't work properly if the data is missing or is invalid. Missing data leads to a change in prediction accuracy. This is very crucial for the prediction framework. It assists the system yield better performance by reducing dimensionality.Figure 3 shows the equation for determining the incomplete quantities.

```
miss_perc="%.2f"%(100*(1-(data[column].dropna().shape[0])/data.shape[0]))
```

Figure 3. Equation for determining the incomplete quantities

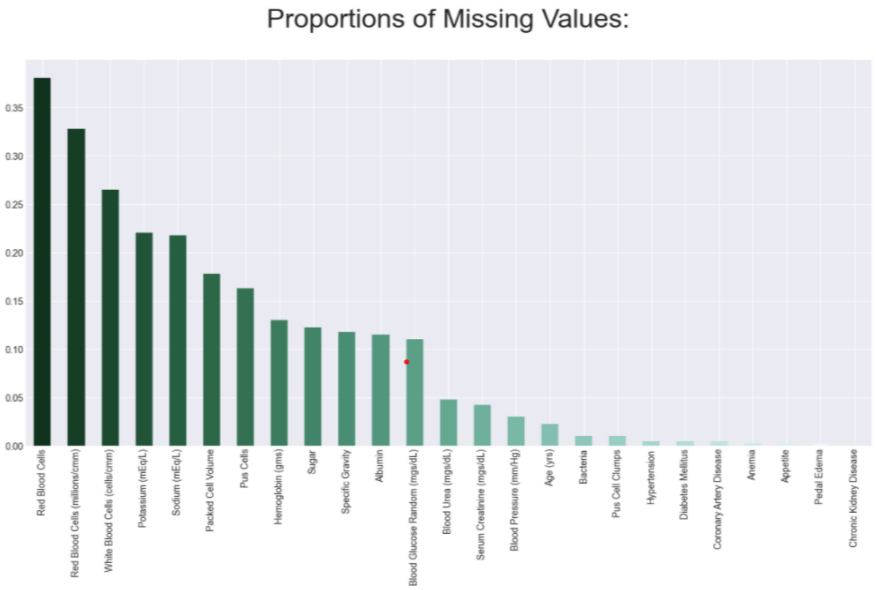


Figure 4. Proportions of missing values in collected data

As you can see in Figure 4, there were missing values in the data that was collected. Using the mean of the observed component is among the most effective approaches to manage incomplete data. The prediction results become better as the data used is genuine. One can also use Regression Imputation to handle the missing values. Regression is a mathematical technique for demonstrating the correlation amongst the dependent and independent parameters This method is simple to understand and appears to be rational. The collected data was taken and processed according to the model. The data collected contains 3 types of values: Nominal Values, Real Values, and Decimal Values. Once the data is collected, Numerical data of the sort 0 and 1 are translated from nominal values. Figure 5 shows steps taken for the collection of this dataset. Both integers and decimal values for various CKD-related properties are included in the ultimate CSV file that was obtained.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

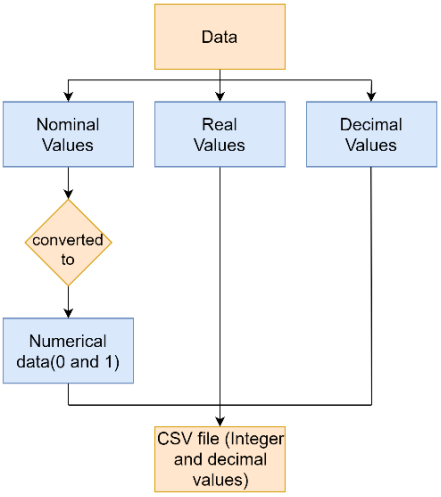


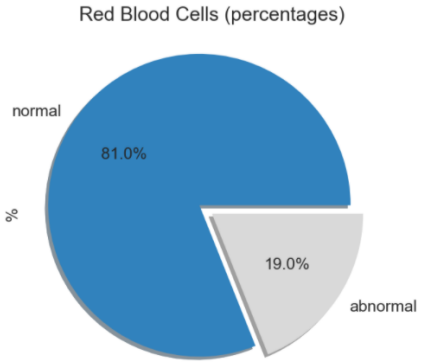
Figure 5. Steps taken for collection of the dataset

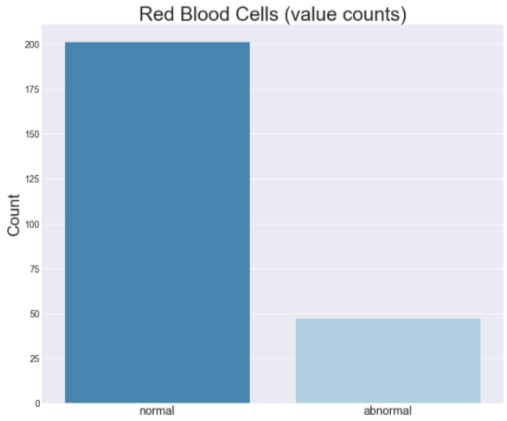
Table 1. Result of the dataset after pre-processing

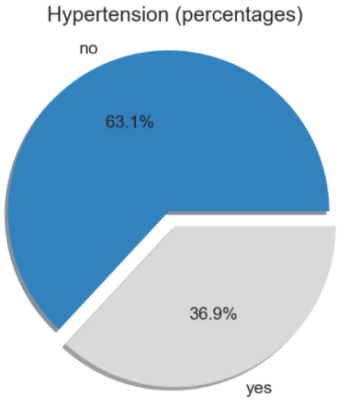
Attributes	Initial Missing Rate (%)	Final Missing Rate (%)
Age	2.25	0
Blood Pressure	3	0
Specific Gravity	11.75	0
Albumin	11.5	0
Sugar	12.25	0
Blood Glucose	11	0
Blood Urea	4.75	0
Serum Creatinine	4.25	0
Sodium	21.75	0
Potassium	22	0
Haemoglobin	13	0
Packed Cell Volume	17.75	0
White Blood Cells	26.50	0
Red Blood Cells	38	0
Pus Cells	16.25	0
Pus Cell Clumps	1	0
Bacteria	1	0
Hypertension	0.5	0
Diabetes Mellitus	0.5	0
Coronary Artery Disease	0.5	0
Appetite	0.25	0
Pedal Edema	0.25	0
Anaemia	0.25	0

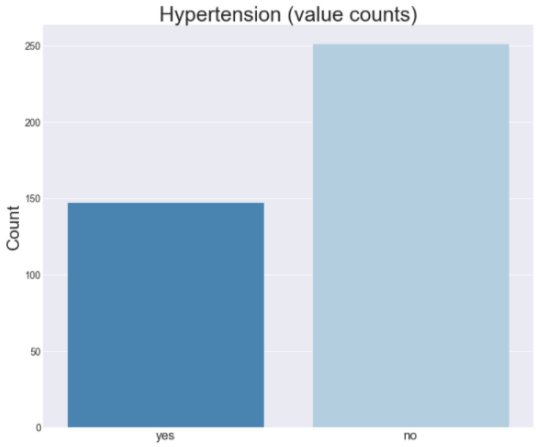
The result of the dataset after Pre-processing is given in Table 1.It is observed that the best results are obtained when missing values are more than 10%. A graph is drawn for the categories of our final data. For example, for Red blood cells, hypertension, Coronary Artery disease, and whether a patient has CKD or not. In Figure 6, analysis of the data collected is shown. Here 'Normal' represents patients who are healthy and 'Abnormal' represents patients who have chronic kidney disease.

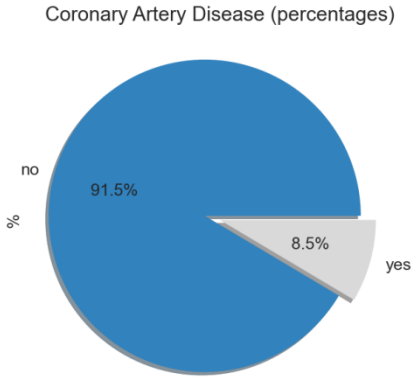
Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

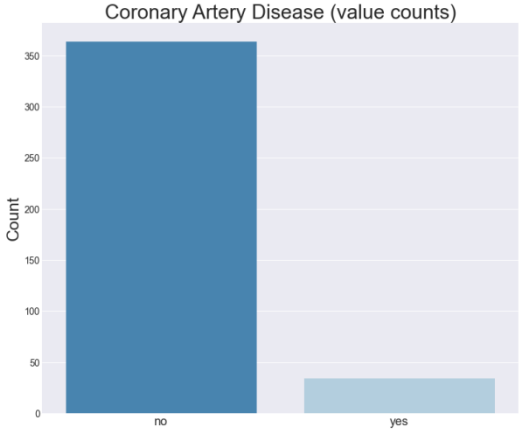












Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

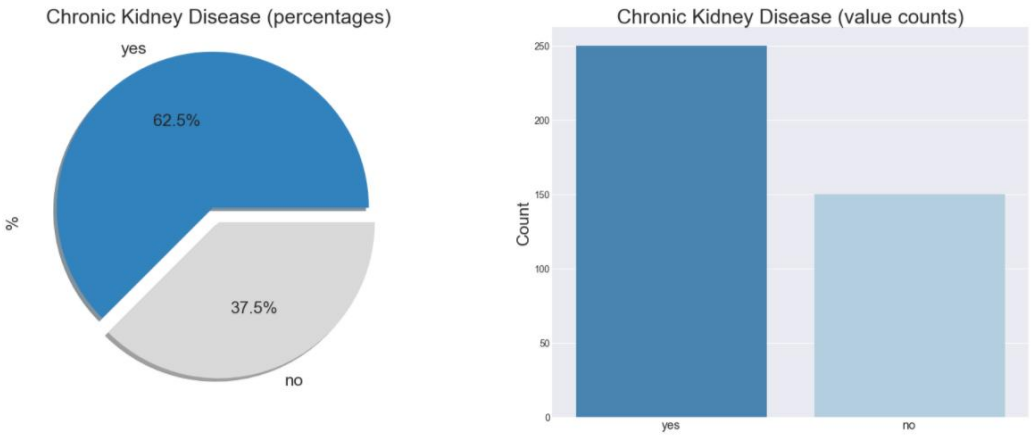


Figure 6. Analysis of data collected

Feature Selection & Model Prediction

As shown in Figure 7, feature selection and data cleaning are considered to have been the initial steps in model design. Feature selection is the method of decreasing the amount of input parameters when creating a predictive model. Few predictive models often contain a wide set of samples which can delay model creation and necessarily require a lot of storage. Furthermore, if input parameters that are unrelated to the target variables were included, the output of certain models will suffer. Thus, Feature Selection is important as it reduces the computational cost of the model that is being developed as well as improves the performance. Data cleaning refers, identifying and correcting the errors which are present in the dataset which can negatively affect the Predictive Model. Some of the errors that may be present in the dataset are duplicated rows, a column that doesn't contain much information, etc. Corrupted data can hinder the model and give incorrect results which may lead to false predictions. Machine learning is a data-driven Artificial Intelligence. If the data is irrelevant or incorrect then the model building is incorrect.

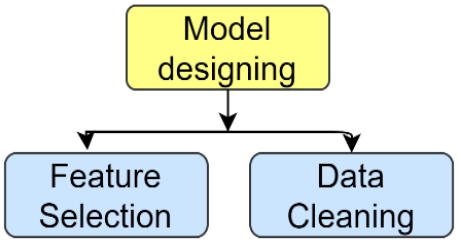


Figure 7. Two types of model designing

The Predictive Model is the idea of creating a system that is worthy of concluding. To make such judgments, a Model comprises numerous algorithms that study those characteristics from a training sample. Our data has numerical as well as categorical variables. Numerical data is quantitative, that is, it is a number and can have any number of possibilities. But categorical data is qualitative. It can have only a limited number of possibilities and it need not be a number. The

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

steps involved in Feature selection are shown in Figure 8. Two Feature selection techniques were used in this analysis to acquire essential attributes. They are Pearson Correlation and ANOVA.

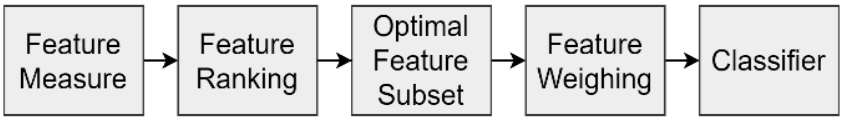


Figure 8. Feature selection

The collection of correlation values among each combination of a dataset's characteristics is organized in a matrix when processing on it. The correlation matrix is the term coined to this matrix. Correlation is indeed a mathematical concept that corresponds to the estimation of the proximity or magnitude of the statistical nature of the association between two entities. There are several ways of determining correlations. Pearson Correlation is by far the most widely used. It calculates the linear relationship between two variables. It estimates the linear association among 2 variables. The Pearson correlation matrix formed here is a 14x14 matrix drawn for the 14 categories that had been taken into consideration.The formula for calculating the Pearson correlation matrix is depicted in Figure 9.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Figure 9. Formula to calculate the Pearson correlation matrix

The Classification Algorithm is applied to the properties that are considered significant and valid after the significant characteristics have been extracted. Univariate Selection is based on the outcome of the univariate statistical analysis in the prototype. The optimal features are considered based on the results of the evaluation. This option relates every feature to the target variable to examine if the feature and thus the target variable has any numerically significant relationship Evaluation of variation is another name for it (ANOVA). While analyzing the relationship between a function and a target variable, all other features are overlooked. Also, every function seems to have a test score associated with it. Eventually, the test results are evaluated, and indeed the features with the highest scores are chosen. For this, K-means clustering is used. By default, there are 8 clusters. Each cluster will have a value based on how well it has performed and much percentage it has learned and been trained. The graph based on the clusters is shown in Figure 10.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

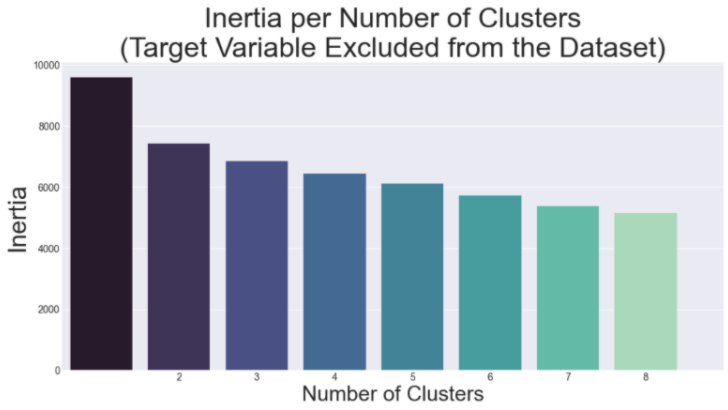


Figure 10. K-means clustering

We may therefore assert that the relative outcome among these two modules leads in contrast to the extracted features within the proposed method against those derived in the previous prototype. Table 2 shows the differences in the extracted functions.

Table 2. Difference between features extracted

Model	Features Extracted
Existing Model	bp, sg, al, pcc, ba, bgr, bu, sod, hemo, pcv, htn, dm, cad, appet, ane
Proposed Model	age, bp, sg, al, su, rbc, pc, pcc, ba, bgr, bu, sc, sod, pot, hemo, pcv, wbcc, rbcc, htn, dm, cad, appet, pe, ane, class

Performance Analysis

Three separate estimation mechanisms were used to determine the efficiency of the techniques. The terms TP, TN, FP, and FN are used in Performance Evaluation. TP is the abbreviation for "True positive." It applies to cases in which the models correctly estimated a positive result. "True Negative" is abbreviated as TN. It applies to situations where the outcome was negative, and the models expected a negative result. FP is an acronym for “False Positive”. It applies to the situation that was anticipated to be positive but turned out to be negative. "False Negative” is abbreviated as FN. It describes a scenario that was expected to be negative but turned out to be positive.

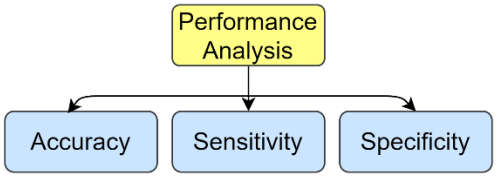


Figure 11. Three types of performance analysis

Figure 11 depicts the three forms of performance evaluation. Accuracy is the proportion of successfully estimated findings (TP+TN) to the ratio of predicted observations. Sensitivity is

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

characterized as the average ratio of true positives accurately detected. Specificity is known as the average ratio of true negatives that are classified accurately.

$$\text{accessibility} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{FP + TN}$$

Figure 12. Formulas for performance analysis

In Figure 12, the equations used for calculating accessibility, sensitivity, and specificity is mentioned as per the characterizations. These were used for analysis of our model. The final scores were varying based on what algorithm was used for the predictive analysis along with neural networks. These results are discussed in the next section.

Results and Discussions

Firstly, the target variable is chosen, which is the classification that whether a person has CKD or not. The confusion matrix is often a common technique for evaluating the classifier as well as determining the classification process's output. However, there are a variety of regular assessment measures for just the classifier's correct and incorrect classification outcomes. Accuracy is the most popular measurement used to assess results.
The confusion matrix was formed, as shown in Figure13, to calculate the values for K-means clusters Vs Target Variable. 194 is Non- CKD patients which have been predicted correctly. 0 false values have been predicted. But out of 150 CKD, 56 has been predicted falsely. So, it is concluded that: 150 is TP, 56 is TN, 196 is FP, and 0 is FN. PCA is a Principal component analysis that is used to find patterns in the dataset. It will give variants for every component. A confusion matrix is created for the combination of PCA and K means against the target variable as shown in Figure 14. From the two matrices, the performance analysis was derived by concatenating all the possible variants for each component. The resultant matrix based on the testing data is demonstrated in Figure 15.

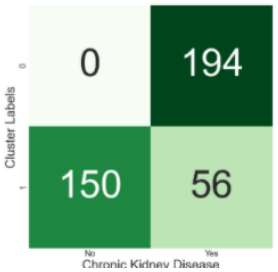


Figure 13. K-means clusters Vs Target variable

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

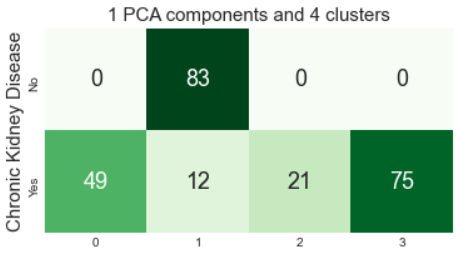


Figure 14. PCA+K-means clusters Vs Target variable

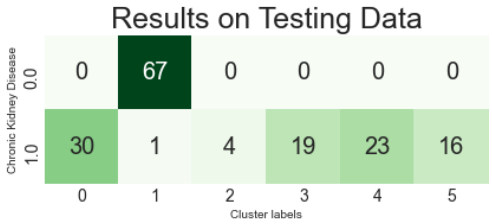


Figure 15. Result on Testing data

To find the accuracy for all the PCA components with other algorithms, various models were used like SVC (Support vector classifier), Nearest neighbor, Random forest, Logistic regression, etc. Below, the charts are drawn for SVM RBF (Support Vector machine- Radial basis function), SVM Poly2 (Support Vector Machine-Polynomial Kernel), Weighted 3NN (Nearest neighbor), Naive Bayes, Logistic Regression, and Random forest. The blue bars give the training accuracy and the purple bars signify the testing accuracy.

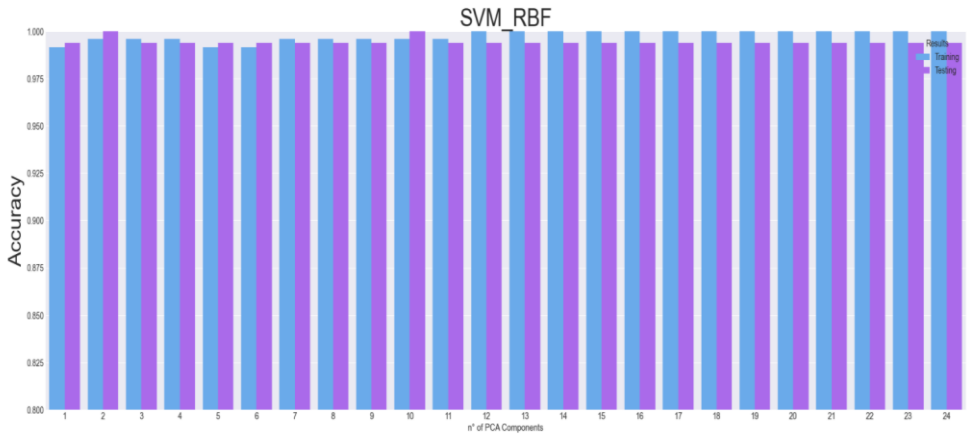


Figure 16. Accuracy of PCA components using SVM RBF

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

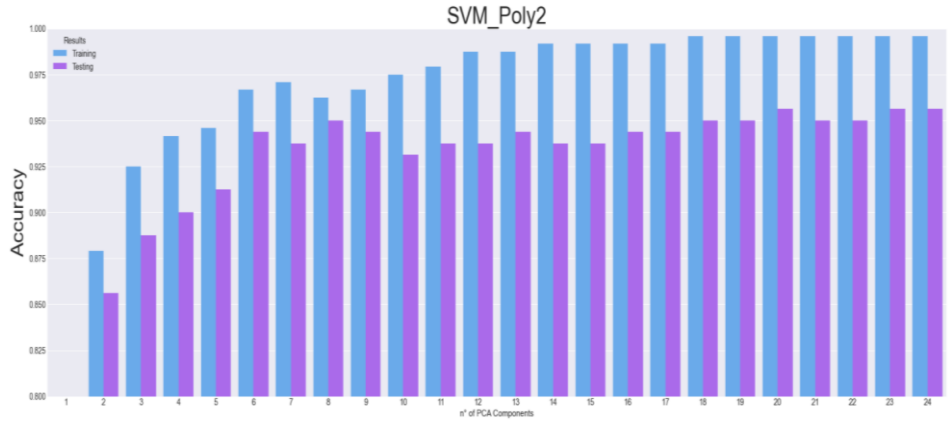


Figure 17. Accuracy of PCA components using SVM Poly2

Figure 18. Accuracy of PCA components using Weighted 3NN

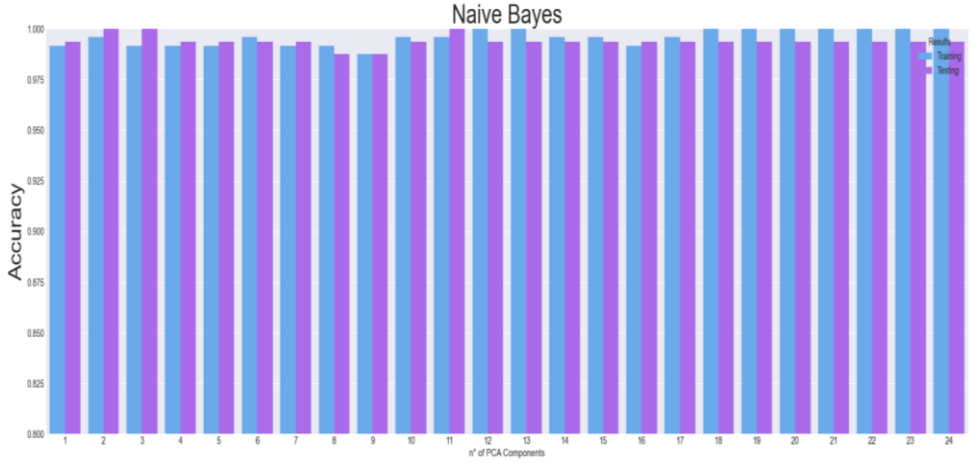


Figure 19. Accuracy of PCA components using Naïve Bayes

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

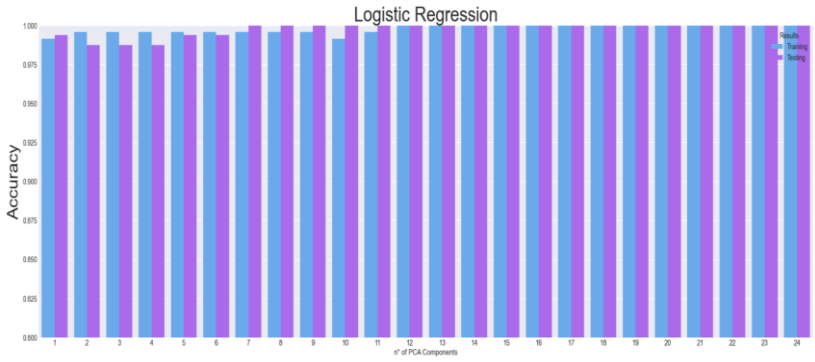


Figure 20. Accuracy of PCA components using Logistic Regression

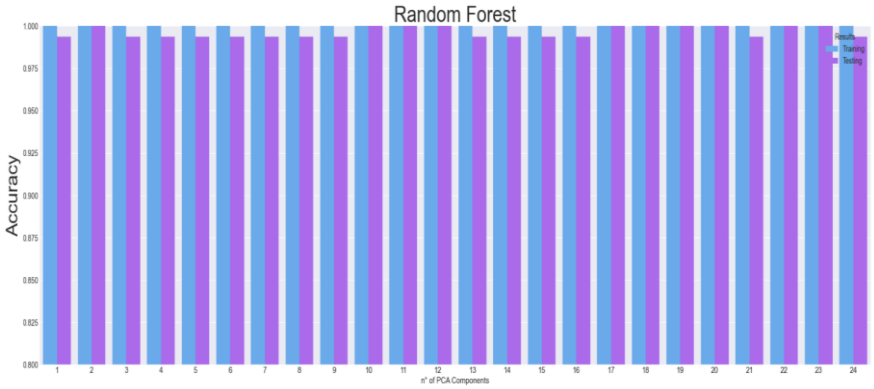


Figure 21. Accuracy of PCA components using Random Forest

Figure 16 to Figure 21, shows the accuracy for various algorithms using neural networks. Similarly, the sensitivity and specificity values based on TP, TN, FP, and FN were calculated. The values altered based on the algorithm used, and the confusion matrix values which has the end results of the training data.

Table 3. Comparison of already existing model with the integrated model

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)
LOG	98.95	98.44	99.80
RF	99.75	99.60	100
Integrated Model	99.83	99.84	99.80
The best average result in this study:			
SVM_RBF	99.54	97.23	99.76
SVM_Poly2	97.47	98.45	100
SVM_Poly3	99.45	100	99.78
Weighted 3NN	98.67	99.92	99.40
Weighted 8NN	98.39	99.94	99.56
Weighted 15NN	98.56	100	99.61
Naïve Bayes	99.52	97.29	98.57
Logistic Regression	99.81	98.73	99.82
Decision Tree	99.44	97.98	98.68
Random Forest	99.75	99.71	100
Integrated Model	99.90	99.89	99.81

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

Table 3 shows a tabular comparison between the results of the already existing model with the proposed integrated model. The formulas of accuracy, sensitivity and specificity are used here in every model here to calculate the outcomes. The results are then converted to percentage values. It is seen that the result of the integrated model is slightly greater than the existing model.

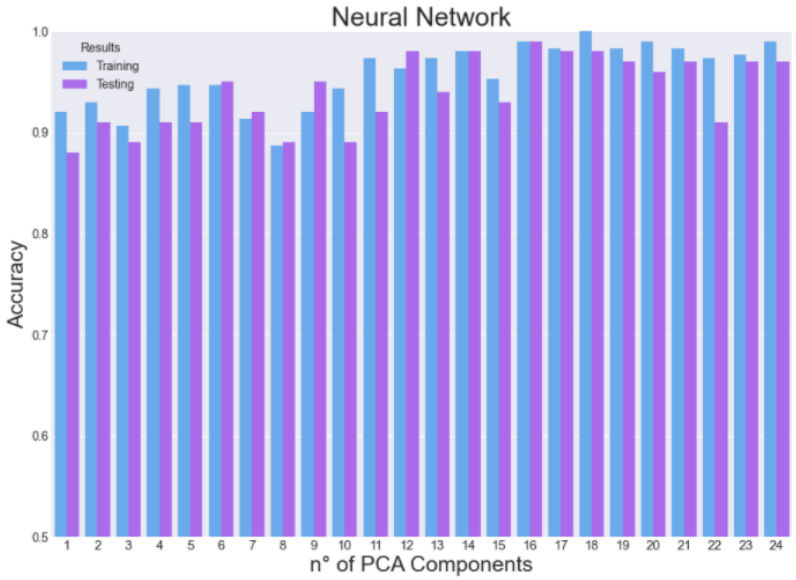


Figure 22. Accuracy of small neural networks

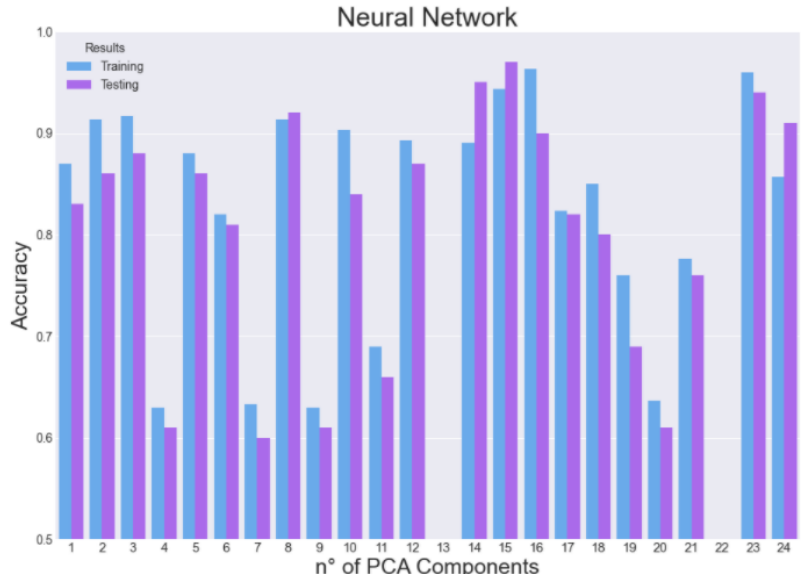


Figure 23. Accuracy of big neural networks

Furthermore, a comparison was done to evaluate the performance using Neural networks. For better results, comparison of results were done for small neural networks as well as big neural

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

networks. All PCA components were considered to plot the graph in Figure 22 and Figure 23. It is observed that small Neural networks give more accuracy than big Neural networks.

Conclusion and Future Work

In this research, we formulated a non-evasive, low-cost, and secure automated prediction method for kidney disease prediction solely on medical records, medical examination, and pathology tests. The sample is pre-processed, and the appropriate aspects from the collection are found using the filter method of feature selection, which includes a univariate selection and correlation matrix. To seek the optimal results for sensitivity, specificity, and accuracy parameters. the output calculations of the classifier of linear kernels were assessed. The future work will focus mainly on scaling down and increasing the accuracy of the algorithm. Other factors can also be considered to make the dataset, and missing values can be avoided or reduced to a minimum amount for better accuracy.

Acknowledgement

We would like to express our heartfelt gratitude to Dr. K. Raja, the Head of the Department and our guide, for his suggestions, guidance, and motivation in completing the research paper smoothly.

References

[1] Mohamed Elhoseny, K. Shankar & J. Uthayakumar, “Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease” published by Springer Nature on 3 July 2019.

[2] Jack Albright, “Forecasting the progression of Alzheimer’s disease using neural networks and a novel preprocessing algorithm” published by Elsevier in 2019.

[3] Vitor A.C. Horta, Ilaria Tiddi, Suzanne Little, Alessandra Mileo a, “Extracting knowledge from Deep Neural Networks through graph analysis” published by Elsevier on 2 March 2021.

[4] Chi Hu, Xiaojun Yu, Qianshan Ding, Zeming Fan, Zhaohui Yuan, Juan Wu, and Linbo Liu, “Cellular-Level Structure Imaging with Micro-optical Coherence Tomography (μOCT) for Kidney Disease Diagnosis” published by IEEE on 9 December 2019.

[5] Mubarik Ahmad, VitriTundjungsari, Dini Widianti, Peny Amalia, UmmiAzizahRachmawati, “Diagnostic Decision Support System of Chronic Kidney Disease Using Support Vector Machine”, published by IEEE on 5 February 2018.

[6] Achim Schilling, Andreas Maier, Richard Gerum, Claus Metzner, Patrick Krauss, “Quantifying the separability of data classes in neural networks”, published by Elsevier on 05 April,2021.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

[7] Arijit Nandi, Nanda Dulal Jana, Swagatam Das, “Improving the Performance of Neural Networks with an Ensemble of Activation Functions”,published in 2020 International Joint Conference on Neural Networks (IJCNN) on 28 September,2020.

[8] Bryan J Moore1, Theodore Berger, Dong Song, “Validation of a Convolutional Neural Network Model for Spike Transformation Using a Generalized Linear Model”,published in 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) on 27 August,2020.

[9] Rajasekaran C, Jayanthi K B, Sudha S and Ramani Kuchelar,“Automated Diagnosis of cardiovascular disease through measurement of intima media thickness using deep neural networks”, published in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) on 07 October,2019.

[10] Heba Mohsen, El-Sayed A. El-Dahshan , El-Sayed M. El-Horbaty, Abdel-Badeeh M. Salem,“Classification using deep learning neural networks for brain tumor”,published by Future Computing and Informatics Journal on 24 December,2017.

[11] Jayson McAllister, Zukui Li, Jinfeng Liu, and Ulrich Simonsmeier, “Erythropoietin Dose Optimization for Anemia in Chronic Kidney Disease Using Recursive Zone Model Predictive Control”, published by IEEE on 23 February,2018.

[12] Anandanadarajah Nishanth, and TharmarajahThiruvaran, “Identifying important attributes for early detection of Chronic Kidney Disease”, published by IEEE Reviews in Biomedical Engineering(Volume:11) on 25 December 2017.

[13] Jiongming qin, lin Chen, Yuhua Liu, chuanjun Liu, changhao Feng, bin Chen,“A Machine Learning Methodology for Diagnosing Chronic kidney disease”, published by IEEE Access(Volume:8) on 30 December 2019.

[14] H. Polat, H.D. Mehr, A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," J. Med. Syst., vol. 41, no. 4, Apr. 2017.

[15] N. R. Hill et al., “Global prevalence of chronic kidney disease - A systematic review and meta-analysis,”Plos One, vol. 11, no. 7, Jul. 2016.

[16] A. M. Cueto-Manzano et al., “Prevalence of chronic kidney disease in an adult population,” Arch. Med. Res., vol. 45, no. 6, pp. 507-513, Aug. 2014.

[17] Xiaonan Zou,Zhewen Tian,Yong Hu,Kaiyuan Shen,“Logistic Regression Model Optimization and Case Analysis”,published in 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT) on 20 January,2020.

[18] Huang Jie1, Wei Yongqing,YiJing and LiuMengdi, “An Improved kNN Based on Class Contribution and Feature Weighting”,published in 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) on 16 April,2018.

[19] Jaime Lynn Speiser , Michael E. Miller, Janet Tooze, Edward Ip,“A comparison of random forest variable selection methods for classification prediction modeling”, published by Elsevier on 23 May,2019.

[20] Min Ding,Zhe Chen, “Wind power prediction based on multiple support vector machines”, published by 2020 39th Chinese Control Conference (CCC) on 09 September,2020.

Annals of R.S.C.B., ISSN: 1583-6258, Vol. 25, Issue 5, 2021, Pages. 5400 - 5420
Received 15 May 2021; Accepted 20 May 2021.

[21] Feng-Jen Yang, "An Implementation of Naive Bayes Classifier",published by 2018 International Conference on Computational Science and Computational Intelligence (CSCI) on 02 January,2020.

[22] Abdullah Al Imran,Md Nur Amin, Fatema TujJohora, "Classification of Chronic Kidney Disease using Logistic Regression, Feedforward Neural Network and Wide & Deep Learning", published in 2018 International Conference on Innovation in Engineering and Technology (ICIET) on 07 March,2019.

[23] Hanyu Zhang, Che-Lun Hung, William Cheng-Chung Chu Ping-Fang Chiu and Chuan Yi Tang, "Chronic Kidney Disease Survival Prediction with Artificial Neural Networks",published in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) on 24 January,2019.

[24] Ravindra B. V, N Sriraam,M.Geetha, "Chronic Kidney Disease Detection using Back Propagation Neural Network Classifier", published by IEEE on 18 March 2019.

[25] Guozhen Chen, Chenguang Ding, Yang Li1, Xiaojun Hu, Xiao Li1, Li Ren, Xiaoming Ding, Puxun Tian, WujunXue, "Prediction of Chronic Kidney Disease using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform", published by IEEE on 18 May 2020.

[26] C. Barbieri et al., "A new machine learning approach for predicting the response to anemia treatment in a large cohort of end-stage renal disease patients undergoing dialysis,"Comput. Biol. Med., vol. 61, pp. 56-61, Jun. 2015.

[27] Shiny Duela, J & Uma Maheswari, P 2018, "Mitigation of DDOS Threat to Service Attainability in Cloud Premises", International Journal of Reasoning-based Intelligent Systems, Vol.10, No.3/4,pp.337-346

[28] Shiny Duela, J & Uma Maheswari, P 2017, "Ensuring Data Security in Cloud Environment through Availability", Journal of Computational and Theoretical Nanoscience, Vol.14, No.9, Pp.4454-4463

[29] Juliet Rani Rajan, Dr.A.ChilambuChelvan, Dr.J.ShinyDuela 2019, "Multi-Class Neural Networks to Predict Lung Cancer", Journal of Medical Systems,43:211, DOI: 10.1007/s10916-019-1355-9,Springer

[30] V. Papademetriou et al., "Chronic kidney disease, basal insulin glargine, and health outcomes in people with dysglycemia: The origin study," Am. J. Med., vol. 130, no. 12, Dec. 2017.