Using Data Mining Techniques for Predicting and Clustering

of Chronic Kidney Disease

Mahima Sharma

Dr. Eyad Haj Said

**Abstract**

Chronic Kidney Disease is a growing global health problem that is associated with high healthcare costs and a high mortality rate. Earlier diagnosis is very important since CKD is irreversible in nature and its symptoms are difficult to detect until the later stages. Data mining techniques can aid physicians in the timely diagnosis of CKD and change the trajectory of the disease's progression. In the first phase of the study, four machine learning models - decision tree, random forest, support vector machine, artificial neural network- were used to establish CKD diagnostic models. The misclassifications made by each model were then evaluated and a hybrid model combining the random forest classifier, SVM classifier, and a naive bayes classifier was developed. The proposed hybrid model performed the best out of all the diagnostic models. In the second phase of the study, a cluster analysis was carried out to identify trends in the positively diagnosed samples and to observe whether the clusters of samples conform to the six clinical stages of CKD. K-Means Clustering and Hierarchical Clustering was implemented in conjunction with PCA, and six clean clusters were observed. However, upon discussion with a specialist, it was concluded that samples in the clusters do not contain enough information to be conclusively labelled as a specific CKD stage.

***Keywords:*** Chronic Kidney Disease, Data mining, Machine Learning, Clustering, Decision tree, Random Forest, Support Vector Machine, Artificial Neural Network, Hybrid Model, K-Means Clustering, Hierarchical Clustering

**Introduction**

Chronic Kidney Disease (CKD) is a steadily growing global health burden with an estimated worldwide prevalence of 13.4%[1](https://www.zotero.org/google-docs/?oTrFNJ). It can be classified into 6 stages of increasing severity- stage 1, 2, 3a, 3b, 4, and 5, where stage 5 is labeled as end-stage kidney disease and implies kidney failure[2](https://www.zotero.org/google-docs/?yXeEsW). The Global Burden of Disease (GBD) study from 2017 estimates that 1.2 million people died from kidney failure in 2017, an increase of 41.5% since 1990[3](https://www.zotero.org/google-docs/?EiaJX5). CKD is also associated with a large economic burden. High-income countries estimatedly spend more than 2–3% of their annual health-care budget on the treatment of end-stage kidney disease in patients who account for under 0.03% of the total population. This situation is worse in most developing nations where nearly one million people die each year because of the unavailability and unaffordability of treatment for end-stage kidney disease[4](https://www.zotero.org/google-docs/?L5IIj8).

CKD is irreversible in nature and eventually progresses to complete loss of renal function. However, it has no observable symptoms in its early stages and is consequently difficult to detect. Machine learning algorithms can be utilized to build a CKD diagnostic model that would allow for the prediction and diagnosis of the disease in its initial stages. Earlier diagnosis would enable patients to receive timely treatments and healthcare advice, and change the trajectory of the disease's progression[5](https://www.zotero.org/google-docs/?e47hRN).

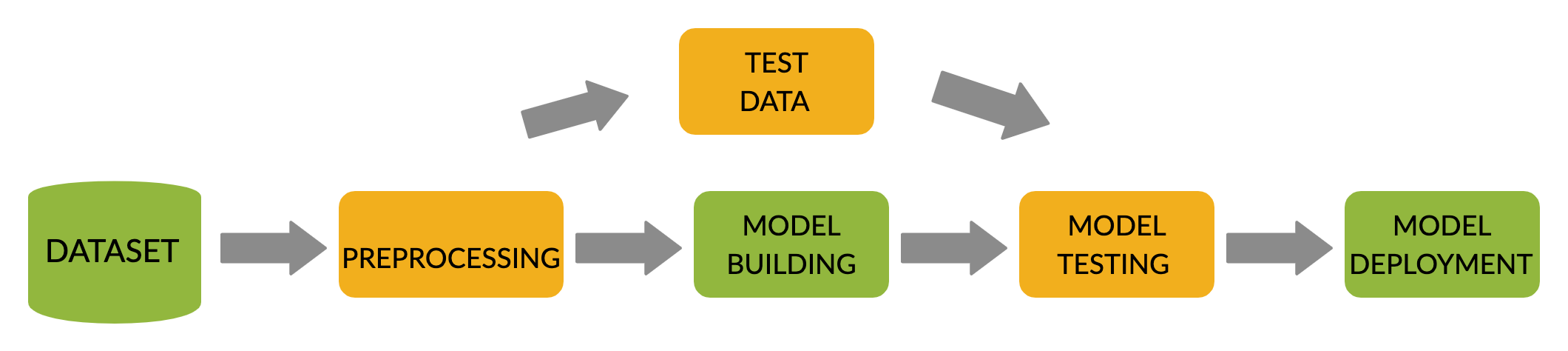
Data mining is the process of extracting meaningful information from a large amount of data, and in recent years, it has found important applications in the health care sector. Large medical datasets are being mined by various machine learning algorithms to predict and classify patients with heart disease[6,7](https://www.zotero.org/google-docs/?e7Yzzg), breast cancer[8](https://www.zotero.org/google-docs/?7ohAKa), diabetes[9](https://www.zotero.org/google-docs/?WdaFmi), and liver cirrhosis[10](https://www.zotero.org/google-docs/?9PgpwE), among other diseases[11–13](https://www.zotero.org/google-docs/?8tWjiC). Clustering algorithms are also being used to identify patterns and find structure in medical datasets pertaining to cardiovascular health[14](https://www.zotero.org/google-docs/?rN6EVV), Irritable Bowel Syndrome[15](https://www.zotero.org/google-docs/?eroYjO), and childhood Asthma[16](https://www.zotero.org/google-docs/?BWrScu). This study is divided into two phases where the first phase focuses on developing CKD diagnostic models and the second phase focuses on a clustering analysis of the individuals with CKD. Four supervised machine learning algorithms- Decision Tree (DT), Random Forest (RF) , Support Vector Machine (SVM), and Artificial Neural Networks (ANN)- were chosen to create CKD diagnostic models. Additionally, a hybrid classifier built from a RF, SVM, and Gaussian Naive Bayes (NB) classifier was also developed for this purpose. The second phase of the study implemented a cluster analysis on the samples with CKD and evaluated the conformity of these clusters to current definitions of CKD severity stages. Two clustering algorithms- K-Means Clustering and Hierarchical Clustering- were used for this purpose.

Within the field of CKD diagnostic models, Polat et al. have applied feature selection methods in conjunction with SVM to diagnose CKD and have achieved an accuracy of 98.5%[14](https://www.zotero.org/google-docs/?PNigBX). Salekin et al. have classified CKD using K-Nearest Neighbors (KNN), Neural Network, and RF and have achieved an accuracy of 99.8% using the RF model[15](https://www.zotero.org/google-docs/?84IBgK). Qin et al. have used LOG, RF, SVM, NB, KNN, Feed-Forward Neural Networks, and an integrated model in their analysis. Their integrated model achieved an accuracy of 100%[16](https://www.zotero.org/google-docs/?4eqPna). One thing to note is that the applications of these CKD diagnostic models is limited to using supervised machine learning to make binary diagnosis predictions. An exception would be the work done by A. Rady et al. where multiclass classification was carried out by first calculating each patient's eGFR to identify the stage of CKD[17](https://www.zotero.org/google-docs/?K4cdhw). However, their research still falls under the umbrella of supervised machine learning. This study focuses on both supervised and unsupervised machine learning models, where the latter has been used to identify clusters in the positively diagnosed samples and see if the clusters can be identified as severity stages of CKD.

The dataset being used in this study was obtained from UCI Machine Learning Repository. The medical data was gathered from Apollo Hospital in Tamilnadu, India over a period of nearly 2 months and donated to the repository in July 2015[18](https://www.zotero.org/google-docs/?tivGIy). The dataset contains 400 patient records and 24 predictive variables (11 numerical variables and 13 categorical variables). The target class has two values- ckd (patient with CKD) and notckd (patient without CKD). The distribution of the target class is slightly skewed since it has 250 patients with CKD and 150 patients without. This dataset is also characterized by a lot of missing values.

**Methods**

The first step in the methodology is to preprocess the data. This may include outlier handling, encoding, scaling, and missing value imputation. The next step is to set aside a portion of the dataset for testing purposes and to build the models by utilizing the remaining dataset. The models then need to be tested using the validation data and their hyperparameters might have to be adjusted based on the performance. Once the models achieve satisfactory performance, they can be deployed[22](https://www.zotero.org/google-docs/?nPPELs). Figure 1 outlines this methodology.



**Figure 1:** Flowchart of Methodology

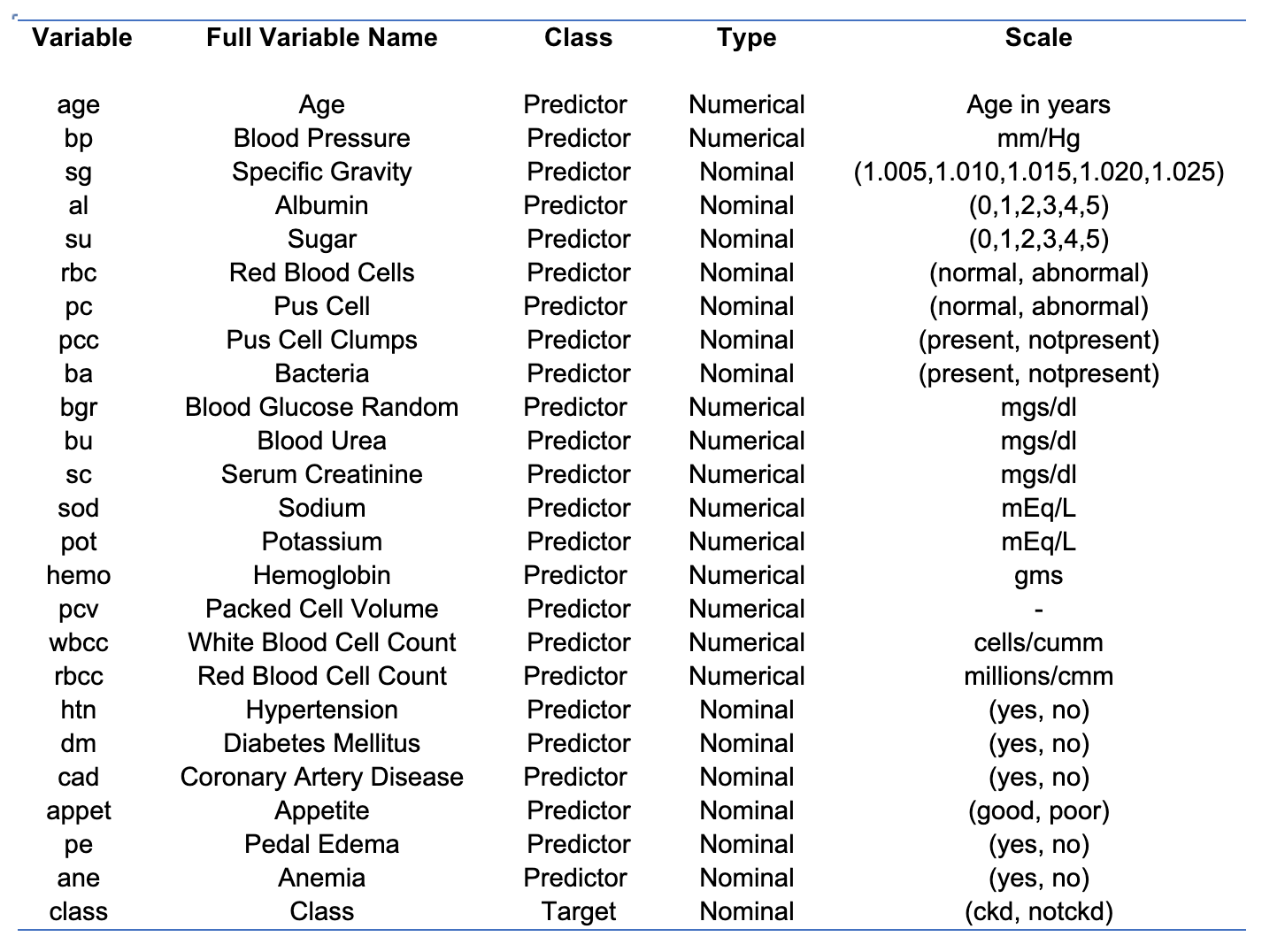
1. **Preprocessing**
   1. **Outlier Handling**

The first preprocessing step executed was outlier handling. An outlier is any data that does not fit in with the other data that is being analyzed. First, the outliers in the CKD dataset were labelled either as a valid outlier or as noise. Valid outliers were interpreted as outlier values that are still medically possible in humans. They were very minimally adjusted, especially those present in clusters, since they might provide important information to the machine learning models. Noise was interpreted as outliers that are not medically possible in humans, indicating some kind of error and rendering the outlier as meaningless. They were replaced with the median value of that predictive variable. The distinction between valid outliers and noise was made through the help of a physician with 20+ years of experience practicing medicine.

* 1. **Encoding**

Since most machine learning models cannot handle nominal variables, the next step was to encode nominal variables in the dataset as numerical values. For the values of rbc and pc, 'normal' and 'abnormal' were encoded as 0 and 1 respectively. the 'notpresent' and 'present' in pcc and ba were encoded as 0 and 1 respectively. For the values of htn, dm, cad, pe, ane, 'no' and 'yes' were encoded as 0 and 1 respectively. The 'good' and 'poor' values in appet were encoded as 0 and 1 respectively.

**Table 1**. Variable Description



* 1. **Feature Scaling**

The next step was Feature Scaling. This involves transforming all the features in the data to fall under a fixed range so that each feature carries equal weight[19](https://www.zotero.org/google-docs/?5FaNSs).

* 1. **Missing Value Imputation**

After scaling, the missing values in the dataset were filled in. For numerical features, the median value of the feature was filled in and for the nominal features, the mode value of the feature was filled in.

1. **Feature Selection**

The next step was feature selection. Feature selection is used to identify and remove features that do not improve the performance or maybe even lower the performance of the machine learning algorithms[23](https://www.zotero.org/google-docs/?IGQtaJ). In this study, three feature selection methods were utilized- Recursive Feature Elimination, Lasso regularization, and Pearson's Correlation. Each of these methods was implemented to select the most important features out of 24 features. A tally was then carried out to count how many times each feature was selected by the three methods and the 8 features with the highest tallies were selected. The selected features are: hemo, sg, al, bgr, htn, dm, appet, and sc. The remaining feature set resulted in a computationally simpler model.

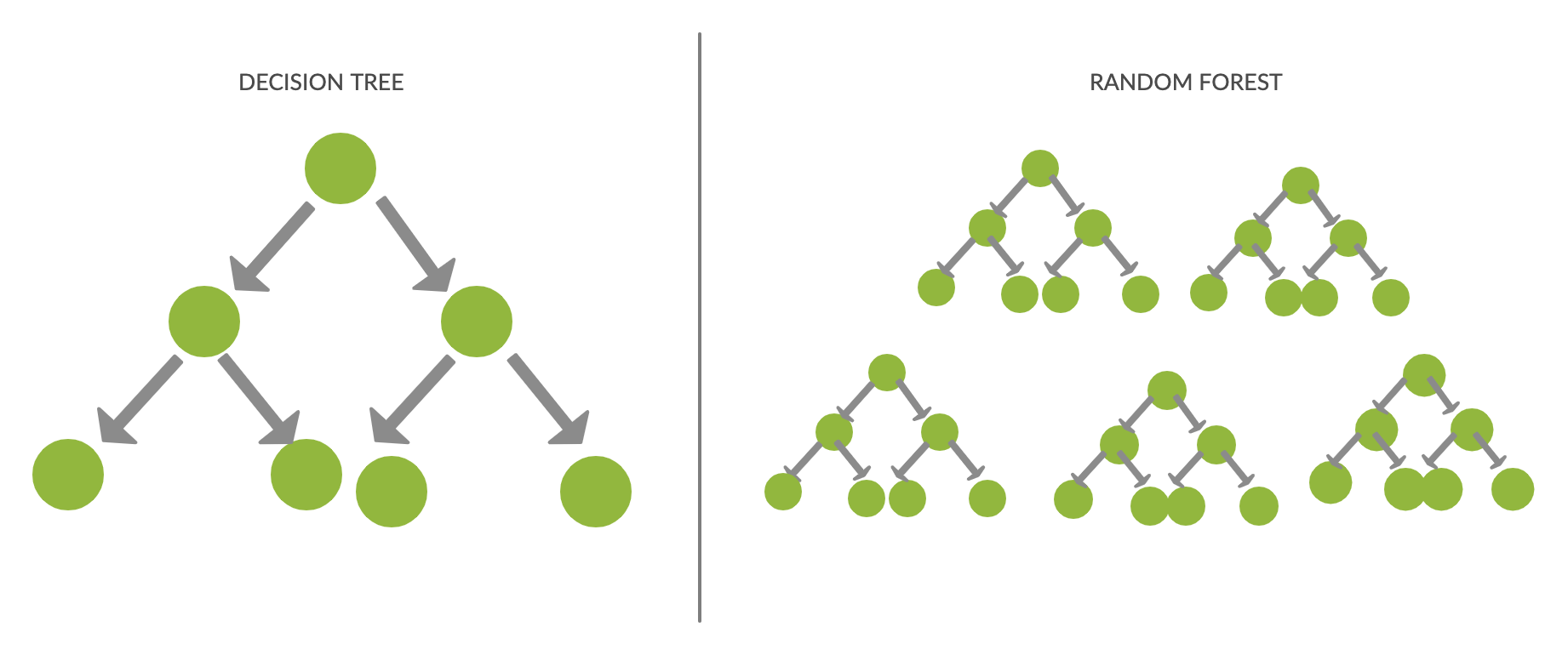
1. **Machine Learning Models**
2. **Prediction and diagnosis of CKD:**

The first phase of this study is to develop a diagnostic model for CKD. In this phase, five different classifiers are examined namely, Decision Tree (DT), Random Forest(RF), Support Vector Machine (SVM), Artificial Neural Networks (ANN), and a hybrid classifier comprising of RF, SVM, and Gaussan Naive Bayes (NB). The process through which these classifiers make diagnostic predictions, can be explained in two phases. They are as follows:

* A training phase where the machine learning model is built by algorithmically learning the relationship between predictive variables and the target variable. A training dataset that includes the target variable is used. In this study, 67% of the CKD dataset was used for the training phase.
* A testing phase where the performance of the model is validated by applying it on a test dataset that excludes the target variable[22](https://www.zotero.org/google-docs/?Quv2Zi) . 33% of the CKD dataset was used for the testing phase.

1. **Decision Tree**

A Decision Tree is a binary tree where predictions are made by traversing from the root node, down the internal nodes, to the leaf nodes. At each of the internal nodes, the data is split into subsets based on the thresholds set by a particular feature. The nodes at which no further splits can be made are called the leaf nodes and they represent a prediction for the target variable[22](https://www.zotero.org/google-docs/?Lcv9ID). Figure 2 represents the structure of the Decision Tree.

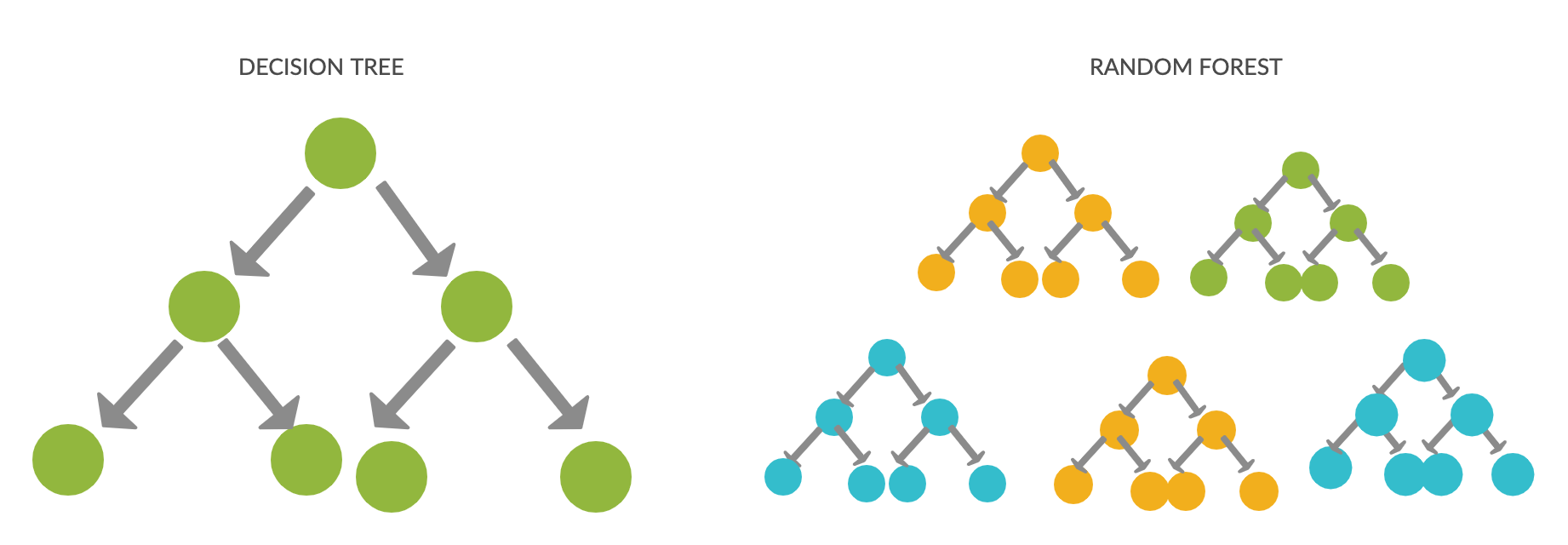


**Figure 2:** Structure of Decision Tree

In this study, hyperparameter tuning was used to choose the following:

criterion = 'gini', max\_depth = 4, min\_samples\_leaf = 1, min\_samples\_split = 2

1. **Random Forest**

Random Forest models consist of an ensemble of many decision trees. The trees are built using a random subset of the training data and the best splits in those trees are calculated from a randomly generated subset of features. Each decision tree makes an output prediction and the output with the most votes is chosen as the model's final prediction. This model was introduced by Leo Breiman in 2001[24](https://www.zotero.org/google-docs/?4iXsnR). Figure 3 outlines the structure of a Random Forest.

**Figure 3:** Structure of Random Forest

In this study, the max\_samples was set as 0.5. Hyperparameter tuning was used to choose the number of trees as 90 and max\_features as 2 out of 8.

1. **Support Vector Machine**

The Support Vector Machine algorithm works by finding the optimal hyperplane in a N-dimensional space (N = number of features) such that the data points of each target class are best separated from those of another target class. This algorithm was developed by Vladimir Vapnik and his colleagues Bernhard Boser and Isabelle Guyon in 1992. It works as follows:

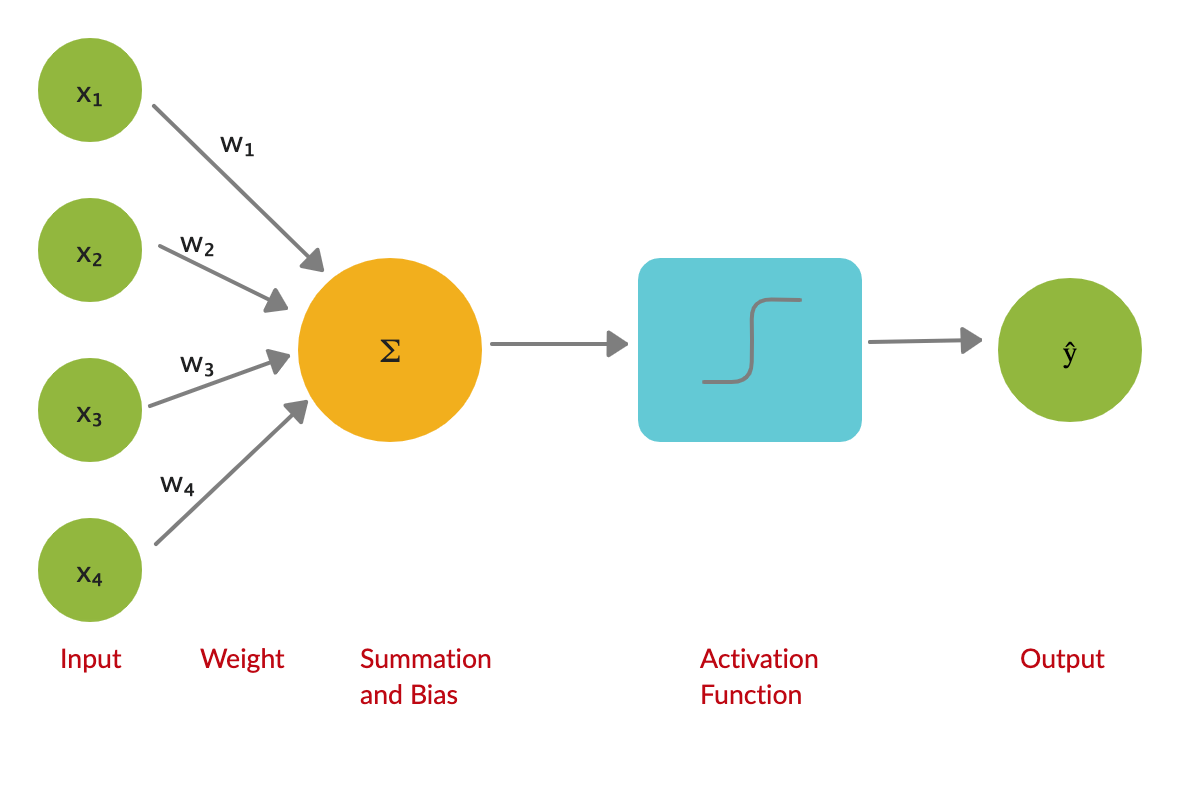
* The points closest to a hyperplane from both the classes are calculated. These are the two support vectors.
* The distance between the hyperplane and the two support vectors is computed. This is called the margin.
* The hyperplane for which the margin is the maximum is called the optimal hyperplane[22](https://www.zotero.org/google-docs/?msVaKe).

In this study, GridSearch was used to find the best set of hyperparameters. C was set as [0.1,1,10,100,1000], ˠ as [1,0.1,0.01, 0.001, 0.0001], and the kernel as ['linear', 'rbf','poly']. The final model was built using C= 100, ˠ = 1, and kernel = 'rbf'.

1. **Artificial Neural Network**

An Artificial Neural Network is a collection of perceptrons (artificial neurons) that are connected in a way that emulates the human brain. An ANN is made up of three different types of perceptron layers- a single input layer, a varying number of hidden layers, and an output layer. The outputs of all the perceptrons in one layer are the inputs to the perceptrons in the next layer.

Each input to a perceptron has a weight associated with it. The perceptron multiplies the inputs and their weights, sums them together with a bias, and then applies the weighted sum to an activation function, where the activation function maps the output to a specific range. These perceptron layers are part of a feedback loop that adjusts the weights of inputs through backpropagation and, in turn, minimizes the loss and maximizes the performance of the model with each iteration[25](https://www.zotero.org/google-docs/?xC1Wy5). Figure 4 outlines a portion of the structure of an ANN.



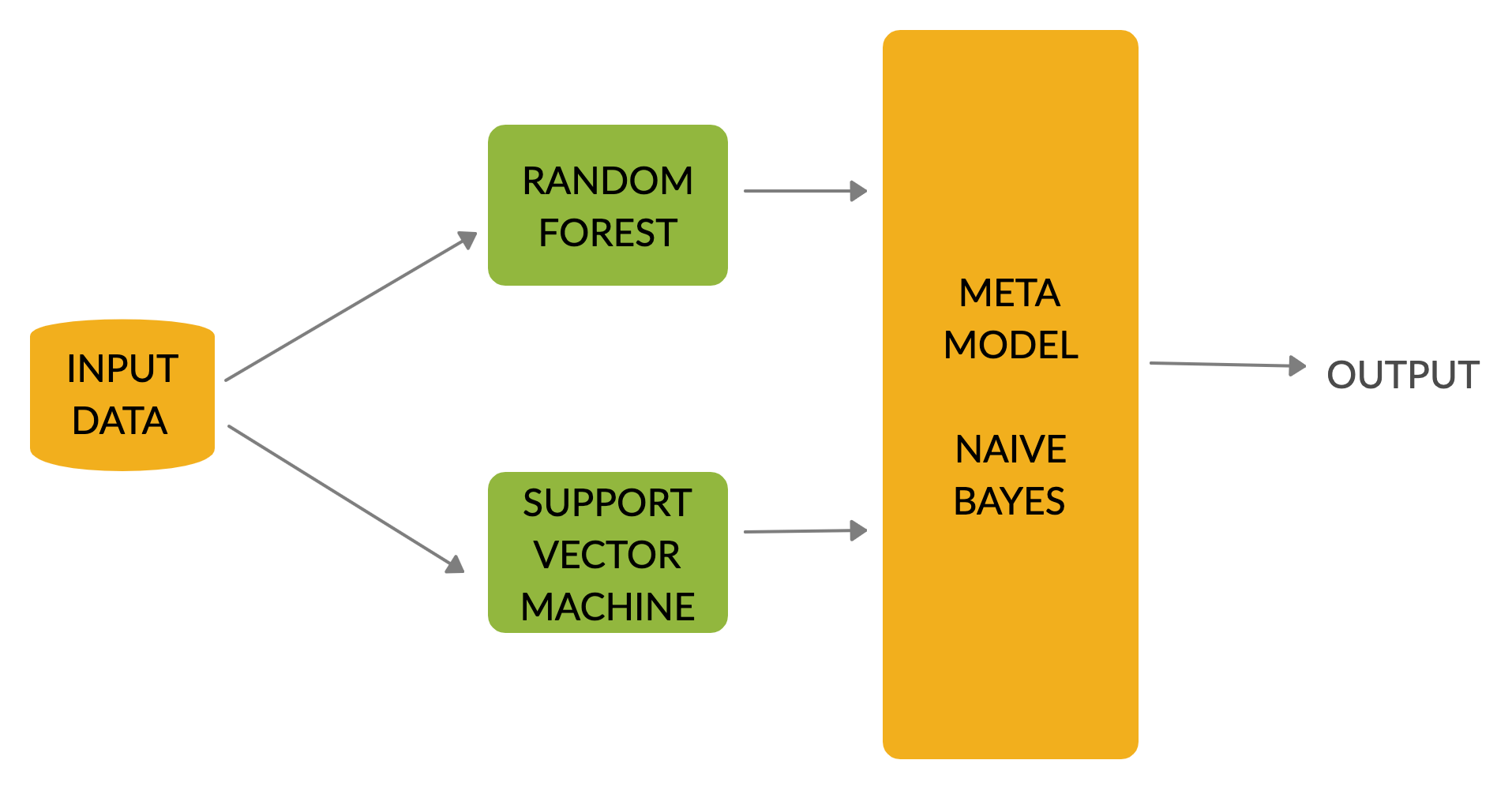
**Figure 4:** Structure of ANN

In this study, an ANN was constructed using Tensorflow and Keras. The model consists of two hidden layers and an output layer. The activation function of the hidden layers is 'relu' and that of the output layer is sigmoid. The optimizer is 'adam' and the loss function is 'binary\_crossentropy'. Although the model runs for 800 iterations, Dropout layers and Early Stopping were used to prevent overfitting of the model.

1. **Hybrid Model**

A hybrid model is established as a type of ensemble classifier that uses the outputs of one classifier as inputs to another classifier. This model structure is based on the hypothesis that combining multiple models together can produce a more powerful model.

In this study, RF and SVM were selected as the base models and NB was selected as the meta-model. This selection was done based on the correlations between the models. Since highly correlated models provide the same information and succeed or fail in diagnosing the same records, selecting all of them as base models would be redundant. A better approach to ensure a more robust final model is to evaluate the misdiagnoses made by individual models and select a mix of high-performing but less-correlated models[26](https://www.zotero.org/google-docs/?nexUK1). Figure 5 represents the proposed model.

****

**Figure 5:** Structure of Hybrid Model

**B. Cluster Analysis of samples with CKD**

The second part of the study was to utilize clustering techniques to observe trends in positively diagnosed samples and see if they can be further clustered into different stages. Two algorithms were used- K-means clustering and Hierarchical Clustering.

1. **K-means Clustering**

K-Means Clustering partitions unlabelled data into K disjoint clusters by selecting K centroids and then assigning each data point to the cluster with the closest centroid. After every data point is assigned, the centroids are recomputed and this process is repeated until no data point changes clusters[27](https://www.zotero.org/google-docs/?q4GhvG).

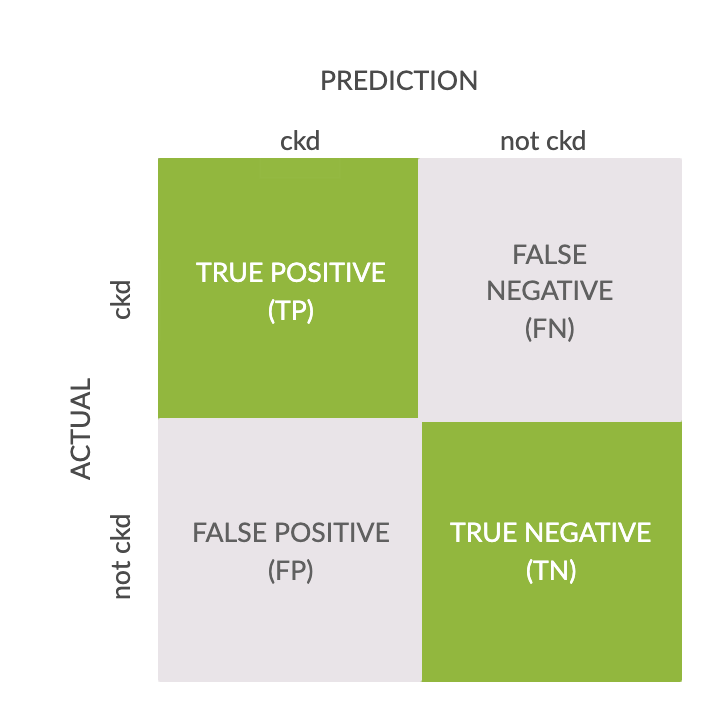
In this study, K-Means Clustering was used in conjunction with Principal Component Analysis (PCA) as a method for visualizing the clusters in two dimensions.

1. **Hierarchical Clustering**

Hierarchical Clustering is a type of cluster analysis that organizes a hierarchy of clusters. It can be divided into two types: Agglomerative and Divisive, where the former was used for this study. In Agglomerative Clustering, each data point is considered as an individual cluster. With each iteration, similar clusters are merged with one another until one or K clusters remain[28](https://www.zotero.org/google-docs/?yfbHOC).

Like with K-Means Clustering, PCA was used for visualizing the clusters in two dimensions.

**D. Evaluation Methods for Diagnostic Models:**

The machine learning models were evaluated using multiple performance metrics. The first metric is the confusion matrix. It is represented as follows:

**Figure 6:** Representation of Confusion Matrix

True positives (TP) are the samples with CKD that were also predicted as having CKD; false negatives (FN) are samples with CKD that were predicted as not having CKD; false positives (FP) are samples without CKD that were predicted as having CKD; and finally, true negatives(TN) are samples without CKD that were predicted as not having CKD.

The TP, FP, FN, TN values in confusion matrices are then plugged into equations to calculate the following performance metrics:

Accuracy: (TP + TN) / (TP + FP + FN + TN)

Precision (False Positive Rate): TP / (TP + FP)

Recall (True Positive Rate): TP / (TP + FN)

F1-Score: (2 \* Precision \* Recall) / Precision + Recall

The performance of the machine learning models can be visualized by using a ROC (Receiver Operating Characteristics) curve. The ROC curve is plotted with true positive rate (TPR) against the false positive rate (FPR) where former is on the y-axis and latter is on the x-axis. The Area Under the Curve (AUC) is a measure of the 2D area under the ROC curve from (0,0) to (1,1). A high AUC indicates well performing machine learning models and a low AUC indicates the opposite[29](https://www.zotero.org/google-docs/?TYPGPk). Therefore, a model with 100% accuracy has an AUC of 1 and a model of 0% accuracy has an AUC of 0.

**Results**

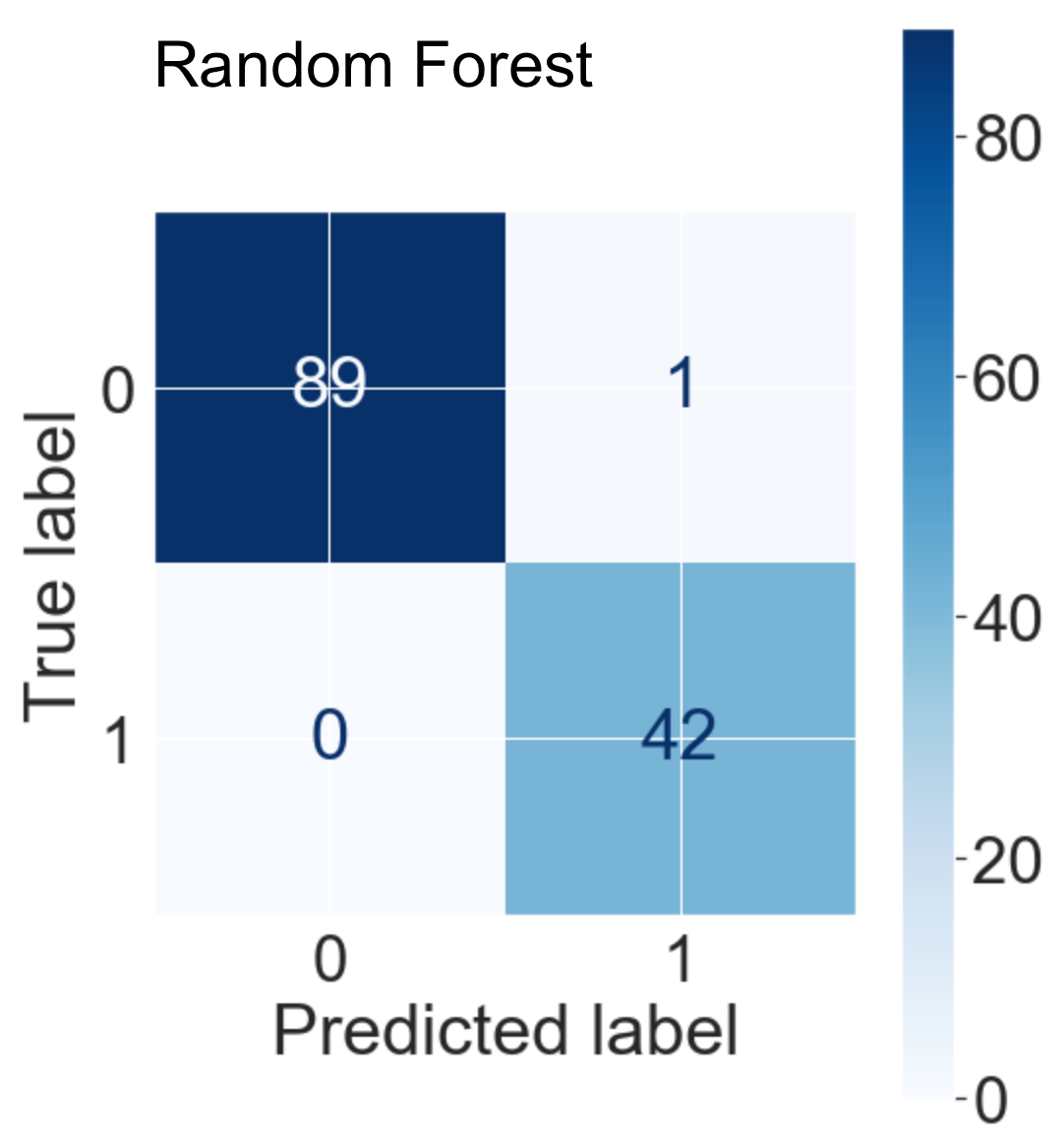
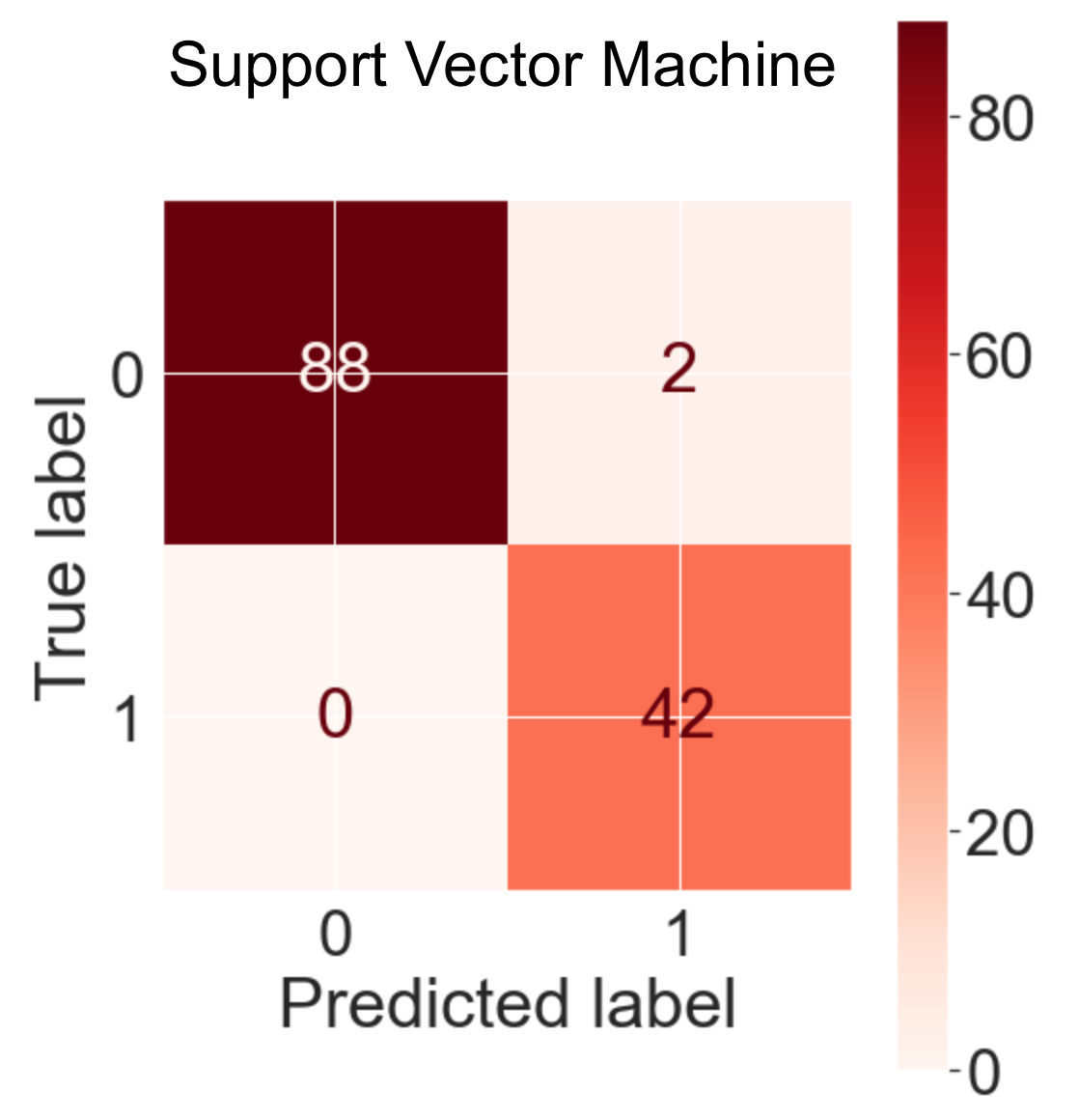
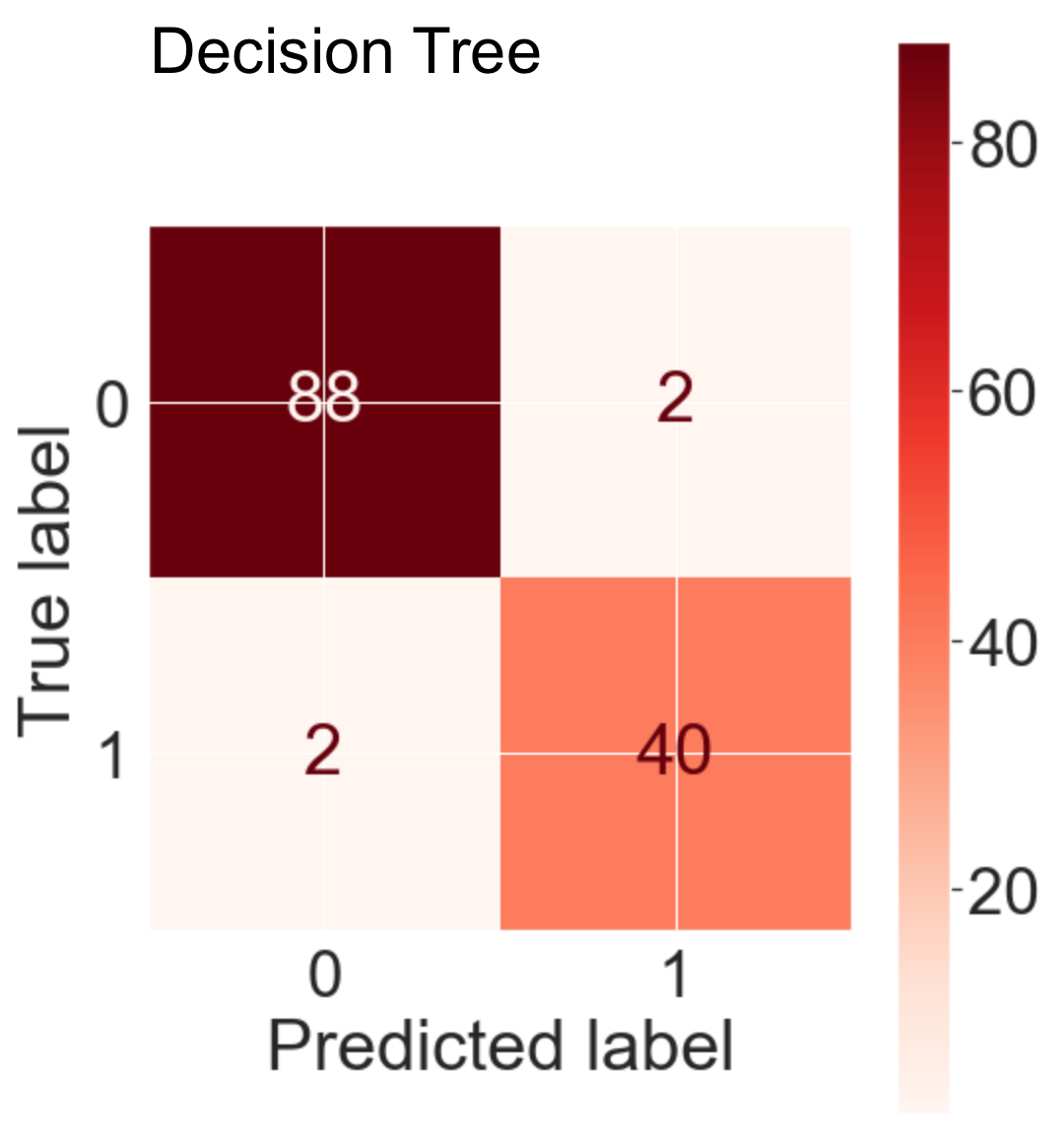
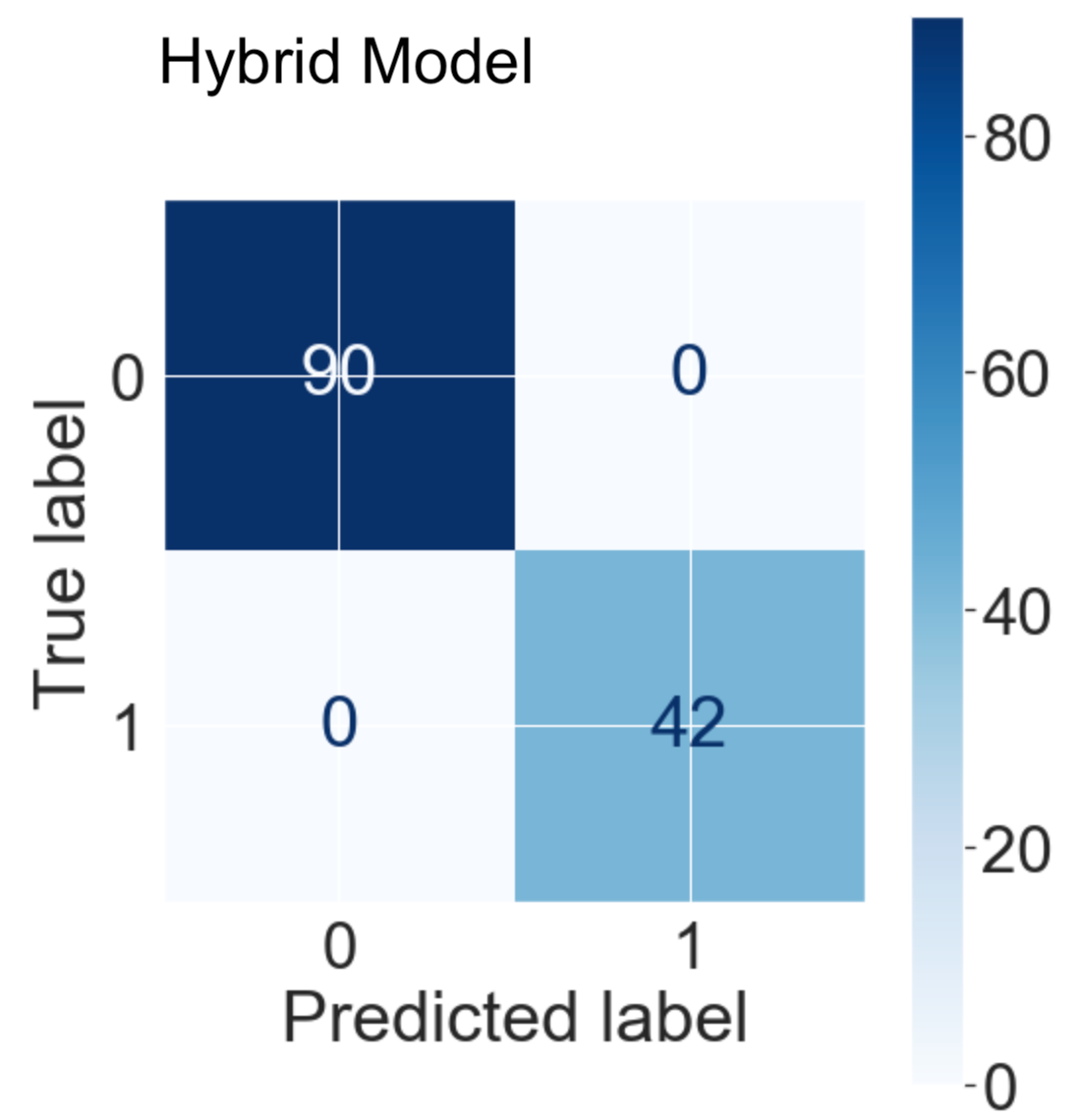
This section is split into two subsections. The first subsection discusses the results from the CKD diagnostic models and the second subsection discusses the results from the clustering algorithms.

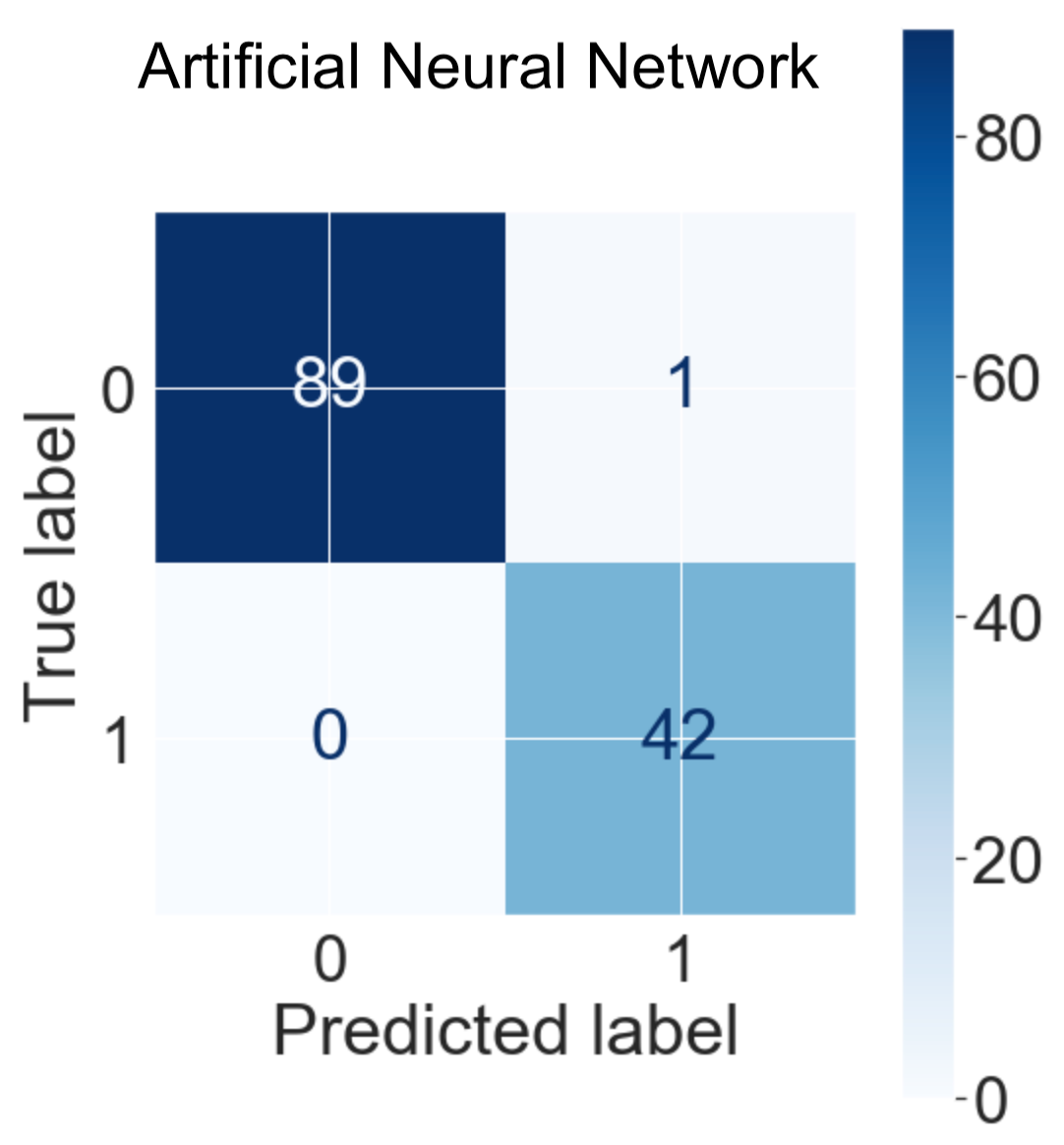
**Results of CKD Diagnostic Models:**

The hybrid classifier gave the most accurate classification, precision, recall, and the highest F-1 score by achieving a score of 100% in each of the metrics. The Random Forest (RF) model and the Artificial Neural Network (ANN) model also performed well as they both have an accuracy of 99% after misdiagnosing 1 record. Support Vector Machine (SVM) achieved an accuracy of 98% after misdiagnosing 2 records while Decision Tree (DT) achieved an accuracy of 96% after misdiagnosing 4 records. The performance of all the CKD diagnostic models are compared in Table 2.

**Table 2** Performance Metrics (in %)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Accuracy** | **Precision**  **(ckd)** | **Precision**  **(not ckd)** | **Recall**  **(ckd)** | **Recall**  **(notckd)** | **F1-Score**  **(ckd)** | **F1-Score**  **(not ckd)** |
| **DT** | 97 | 98 | 95 | 98 | 95 | 98 | 95 |
| **RF** | 99 | 100 | 98 | 99 | 100 | 99 | 99 |
| **SVM** | 98 | 100 | 95 | 98 | 100 | 99 | 98 |
| **ANN** | 99 | 100 | 98 | 99 | 100 | 99 | 99 |
| **Hybrid Classifier** | 100 | 100 | 100 | 100 | 100 | 100 | 100 |



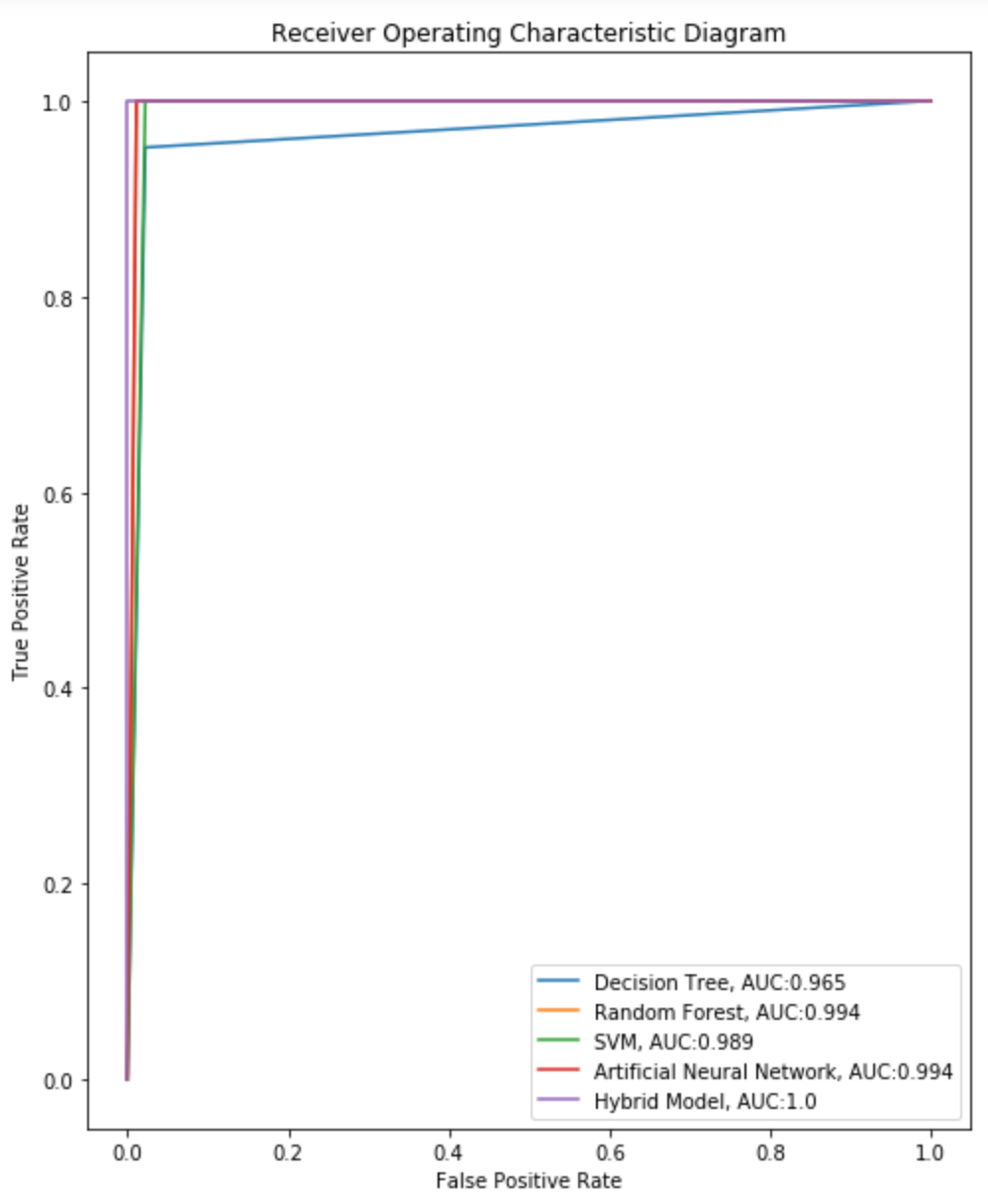


**Figure 7:** Confusion Matrices of CKD Diagnostic Models

The AUC values of the models are represented in Table 3. The hybrid classifier performed the best among all the models with an AUC of 1.

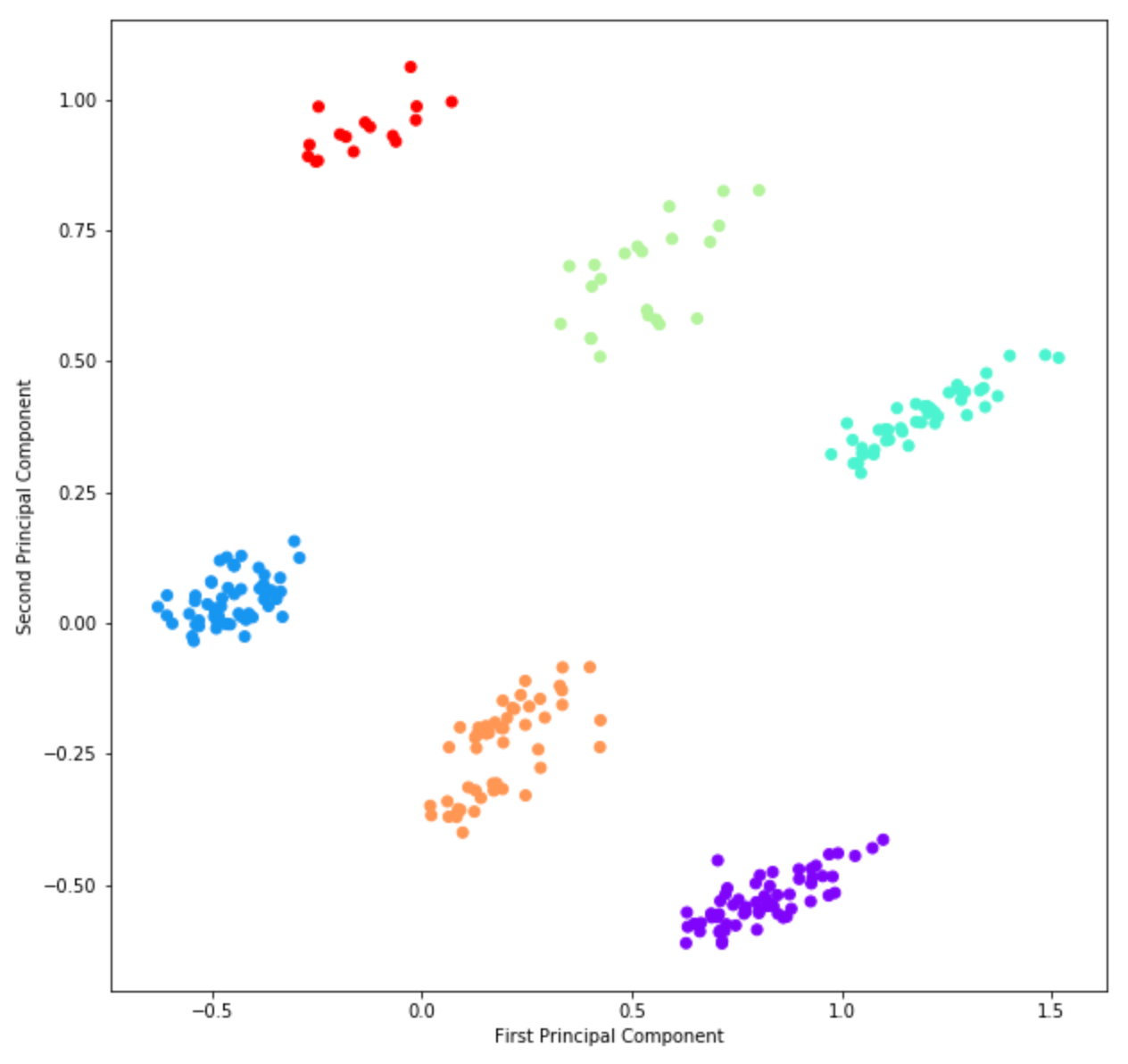
**Table 3** AUC values

|  |  |
| --- | --- |
| **Algorithm** | **AUC** |
| **DT** | 0.965 |
| **RF** | 0.994 |
| **SVM** | 0.988 |
| **ANN** | 0.994 |
| **Hybrid Classifier** | 1.000 |

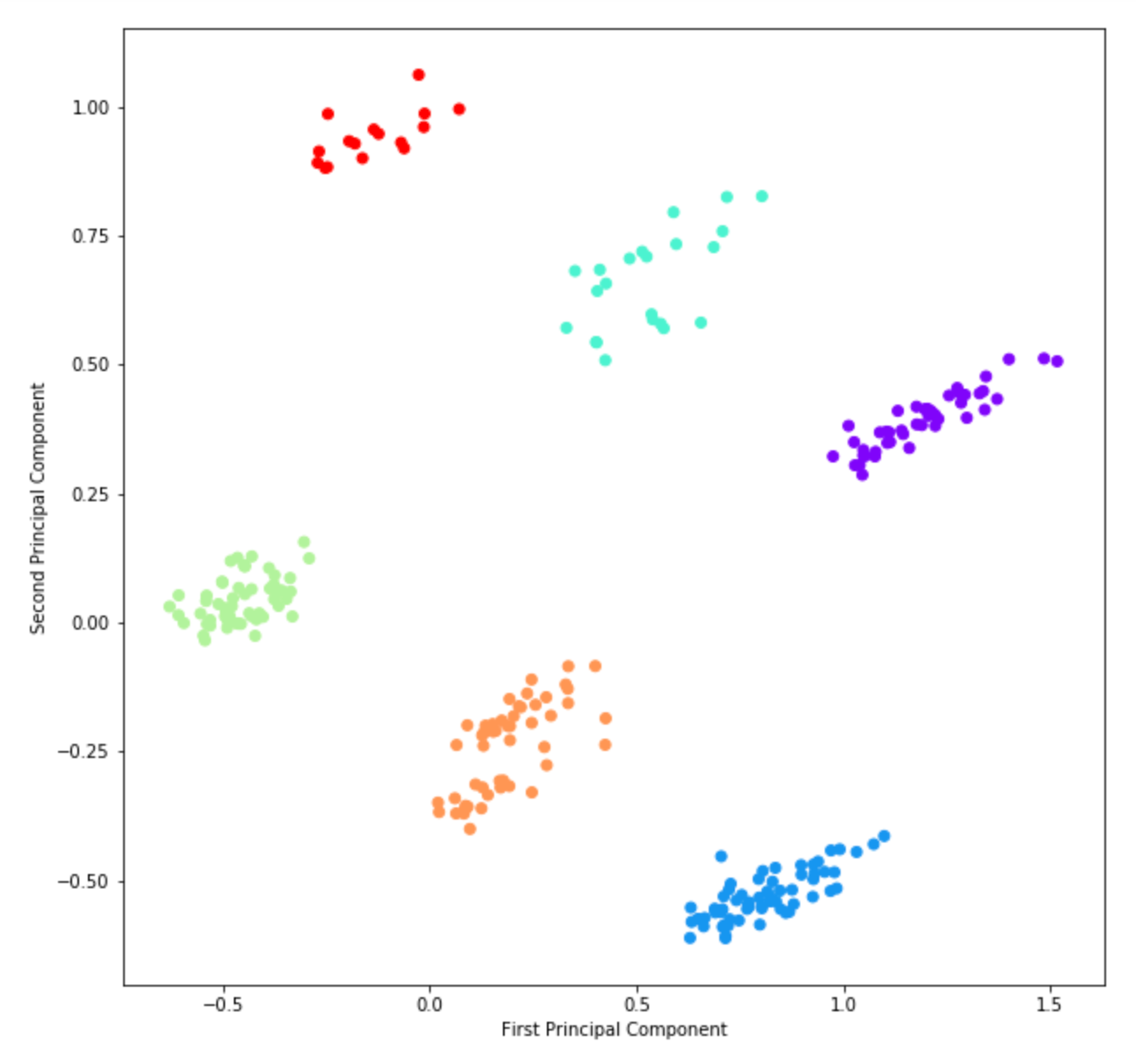
**Figure 8:** ROC curves of each ML Model

**Results of Clustering Algorithms**

After using K-Means clustering in conjunction with Principal Component Analysis (PCA), 6 cleanly-separated clusters were observed. These clusters are shown in Figure 9. These results were mirrored by the Hierarchical Clustering algorithm, as shown in Figure 10.



**Figure 9**: Results of K-Means Clustering



**Figure 10:** Results of Hierarchical Clustering

After getting these clusters, the samples in each cluster were evaluated for trends. It was observed that some clusters had patients with higher levels of serum creatinine, higher incidences of diabetes and hypertension, and worse appetite than other clusters. These records were shown to a nephrologist who suggested that although the trend might be indicative of worsening CKD, it is difficult to label the clusters as different stages of CKD without calculating the Glomerular Filtration Rate (GFR) of the patients. One of the most accurate methods for calculating the GFR is is the Modification of Diet in Renal Disease (MDRD) and it is given by the equation below:

Since the current dataset does not contain information about samples' gender and race, additional information is required to test this hypothesis further. At the current stage of this study, it is not possible to calculate the GFR and definitively identify the stages of CKD.

**Discussion and Conclusion**

The first part of the study suggests that data mining techniques can be utilized to diagnose Chronic Kidney Disease. All of the machine learning models in this study performed well but the hybrid classifier performed the best with an accuracy of 100% and an AUC value of 1.0. This is to be expected since hybrid classifiers generally perform better than single learners[26](https://www.zotero.org/google-docs/?mgbEEp). The performance of the hybrid classifier was followed closely by Artificial Neural Network and Random Forest models with an accuracy of 99% and an AUC value of 0.994, Support Vector Machine with an accuracy of 98% and an AUC of 0.988, and Decision Tree with an accuracy of 96% and an AUC of 0.965. Although the performance of the models did not vary much when considering numerical measures, it varies a lot more when the biological implications are considered. The cost of a diagnosis error on human life can be high, especially if a sample with CKD is misdiagnosed as not having it. Something worth mentioning is that the biological information imported into the machine learning models are important clinical clues in the diagnosis of CKD30. Our confidence in the reliability of the models' predictions is supported by the knowledge that the appropriate biological markers were mined while diagnosing CKD.

Although the results of the diagnostic models are optimistic, an important limitation should be highlighted. The current dataset has a small sample size, containing only 400 samples. Given its size, it might not be representative of all cases of CKD. Therefore, it is unclear whether the performance of the models can be generalized over a larger dataset. Increasing the sample size is an important next-step to evaluating and optimizing these models further.

The results from the second part of the study suggest that the CKD samples can be clustered. However, it is currently not clear whether the clusters simply represent patterns in the dataset or if they contain more information, like the severity stages of CKD. As the duration of this research was limited to eight weeks, the work related to clustering remains an area where further analysis can be done in the future. Additionally, if more data is gathered to mitigate the current limitation in sample size, it might also be beneficial to include race and gender as potential attributes for the new samples. This would allow the calculation of the GFR and make way for more work to be done in regards to CKD stages.

Bibliography

[(1) Coresh, J. Update on the Burden of CKD. *J. Am. Soc. Nephrol.* **2017**, *28* (4), 1020–1022. https://doi.org/10.1681/ASN.2016121374.](https://www.zotero.org/google-docs/?nxPdeA)

[(2) Thomas, R.; Kanso, A.; Sedor, J. R. Chronic Kidney Disease and Its Complications. *Prim. Care* **2008**, *35* (2), 329–vii. https://doi.org/10.1016/j.pop.2008.01.008.](https://www.zotero.org/google-docs/?nxPdeA)

[(3) Bikbov, B.; Purcell, C. A.; Levey, A. S.; Smith, M.; Abdoli, A.; Abebe, M.; Adebayo, O. M.; Afarideh, M.; Agarwal, S. K.; Agudelo-Botero, M.; Ahmadian, E.; Al-Aly, Z.; Alipour, V.; Almasi-Hashiani, A.; Al-Raddadi, R. M.; Alvis-Guzman, N.; Amini, S.; Andrei, T.; Andrei, C. L.; Andualem, Z.; Anjomshoa, M.; Arabloo, J.; Ashagre, A. F.; Asmelash, D.; Ataro, Z.; Atout, M. M. W.; Ayanore, M. A.; Badawi, A.; Bakhtiari, A.; Ballew, S. H.; Balouchi, A.; Banach, M.; Barquera, S.; Basu, S.; Bayih, M. T.; Bedi, N.; Bello, A. K.; Bensenor, I. M.; Bijani, A.; Boloor, A.; Borzì, A. M.; Cámera, L. A.; Carrero, J. J.; Carvalho, F.; Castro, F.; Catalá-López, F.; Chang, A. R.; Chin, K. L.; Chung, S.-C.; Cirillo, M.; Cousin, E.; Dandona, L.; Dandona, R.; Daryani, A.; Gupta, R. D.; Demeke, F. M.; Demoz, G. T.; Desta, D. M.; Do, H. P.; Duncan, B. B.; Eftekhari, A.; Esteghamati, A.; Fatima, S. S.; Fernandes, J. C.; Fernandes, E.; Fischer, F.; Freitas, M.; Gad, M. M.; Gebremeskel, G. G.; Gebresillassie, B. M.; Geta, B.; Ghafourifard, M.; Ghajar, A.; Ghith, N.; Gill, P. S.; Ginawi, I. A.; Gupta, R.; Hafezi-Nejad, N.; Haj-Mirzaian, A.; Haj-Mirzaian, A.; Hariyani, N.; Hasan, M.; Hasankhani, M.; Hasanzadeh, A.; Hassen, H. Y.; Hay, S. I.; Heidari, B.; Herteliu, C.; Hoang, C. L.; Hosseini, M.; Hostiuc, M.; Irvani, S. S. N.; Islam, S. M. S.; Balalami, N. J.; James, S. L.; Jassal, S. K.; Jha, V.; Jonas, J. B.; Joukar, F.; Jozwiak, J. J.; Kabir, A.; Kahsay, A.; Kasaeian, A.; Kassa, T. D.; Kassaye, H. G.; Khader, Y. S.; Khalilov, R.; Khan, E. A.; Khan, M. S.; Khang, Y.-H.; Kisa, A.; Kovesdy, C. P.; Defo, B. K.; Kumar, G. A.; Larsson, A. O.; Lim, L.-L.; Lopez, A. D.; Lotufo, P. A.; Majeed, A.; Malekzadeh, R.; März, W.; Masaka, A.; Meheretu, H. A. A.; Miazgowski, T.; Mirica, A.; Mirrakhimov, E. M.; Mithra, P.; Moazen, B.; Mohammad, D. K.; Mohammadpourhodki, R.; Mohammed, S.; Mokdad, A. H.; Morales, L.; Velasquez, I. M.; Mousavi, S. M.; Mukhopadhyay, S.; Nachega, J. B.; Nadkarni, G. N.; Nansseu, J. R.; Natarajan, G.; Nazari, J.; Neal, B.; Negoi, R. I.; Nguyen, C. T.; Nikbakhsh, R.; Noubiap, J. J.; Nowak, C.; Olagunju, A. T.; Ortiz, A.; Owolabi, M. O.; Palladino, R.; Pathak, M.; Poustchi, H.; Prakash, S.; Prasad, N.; Rafiei, A.; Raju, S. B.; Ramezanzadeh, K.; Rawaf, S.; Rawaf, D. L.; Rawal, L.; Reiner, R. C.; Rezapour, A.; Ribeiro, D. C.; Roever, L.; Rothenbacher, D.; Rwegerera, G. M.; Saadatagah, S.; Safari, S.; Sahle, B. W.; Salem, H.; Sanabria, J.; Santos, I. S.; Sarveazad, A.; Sawhney, M.; Schaeffner, E.; Schmidt, M. I.; Schutte, A. E.; Sepanlou, S. G.; Shaikh, M. A.; Sharafi, Z.; Sharif, M.; Sharifi, A.; Silva, D. A. S.; Singh, J. A.; Singh, N. P.; Sisay, M. M. M.; Soheili, A.; Sutradhar, I.; Teklehaimanot, B. F.; Tesfay, B. etsay; Teshome, G. F.; Thakur, J. S.; Tonelli, M.; Tran, K. B.; Tran, B. X.; Ngoc, C. T.; Ullah, I.; Valdez, P. R.; Varughese, S.; Vos, T.; Vu, L. G.; Waheed, Y.; Werdecker, A.; Wolde, H. F.; Wondmieneh, A. B.; Hanson, S. W.; Yamada, T.; Yeshaw, Y.; Yonemoto, N.; Yusefzadeh, H.; Zaidi, Z.; Zaki, L.; Zaman, S. B.; Zamora, N.; Zarghi, A.; Zewdie, K. A.; Ärnlöv, J.; Coresh, J.; Perico, N.; Remuzzi, G.; Murray, C. J. L.; Vos, T. Global, Regional, and National Burden of Chronic Kidney Disease, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet* **2020**, *395* (10225), 709–733. https://doi.org/10.1016/S0140-6736(20)30045-3.](https://www.zotero.org/google-docs/?nxPdeA)

[(4) WHO | The global burden of kidney disease and the sustainable development goals http://www.who.int/bulletin/volumes/96/6/17-206441/en/ (accessed Jun 15, 2020). https://doi.org/10.2471/BLT.17.206441.](https://www.zotero.org/google-docs/?nxPdeA)

[(5) James, M. T.; Hemmelgarn, B. R.; Tonelli, M. Early Recognition and Prevention of Chronic Kidney Disease. *The Lancet* **2010**, *375* (9722), 1296–1309. https://doi.org/10.1016/S0140-6736(09)62004-3.](https://www.zotero.org/google-docs/?nxPdeA)

[(6) Desai, S. D.; Dessai, I. F.; Kulkarni, L. *Intelligent Heart Disease Prediction System Using Probabilistic Neural Network*.](https://www.zotero.org/google-docs/?nxPdeA)

[(7) Masetic, Z.; Subasi, A. Congestive Heart Failure Detection Using Random Forest Classifier. *Comput. Methods Programs Biomed.* **2016**, *130*, 54–64. https://doi.org/10.1016/j.cmpb.2016.03.020.](https://www.zotero.org/google-docs/?nxPdeA)

[(8) Patrício, M.; Pereira, J.; Crisóstomo, J.; Matafome, P.; Gomes, M.; Seiça, R.; Caramelo, F. Using Resistin, Glucose, Age and BMI to Predict the Presence of Breast Cancer. *BMC Cancer* **2018**, *18* (1), 29. https://doi.org/10.1186/s12885-017-3877-1.](https://www.zotero.org/google-docs/?nxPdeA)

[(9) Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting Diabetes Mellitus With Machine Learning Techniques. *Front. Genet.* **2018**, *9*. https://doi.org/10.3389/fgene.2018.00515.](https://www.zotero.org/google-docs/?nxPdeA)

[(10) Cao, Y.; Hu, Z.-D.; Liu, X.-F.; Deng, A.-M.; Hu, C.-J. An MLP Classifier for Prediction of HBV-Induced Liver Cirrhosis Using Routinely Available Clinical Parameters. *Dis. Markers* **2013**, *35* (6), 653–660. https://doi.org/10.1155/2013/127962.](https://www.zotero.org/google-docs/?nxPdeA)

[(11) Magnin, B.; Mesrob, L.; Kinkingnéhun, S.; Pélégrini-Issac, M.; Colliot, O.; Sarazin, M.; Dubois, B.; Lehéricy, S.; Benali, H. Support Vector Machine-Based Classification of Alzheimer’s Disease from Whole-Brain Anatomical MRI. *Neuroradiology* **2009**, *51* (2), 73–83. https://doi.org/10.1007/s00234-008-0463-x.](https://www.zotero.org/google-docs/?nxPdeA)

[(12) Miljkovic, D.; Aleksovski, D.; Podpečan, V.; Lavrač, N.; Malle, B.; Holzinger, A. Machine Learning and Data Mining Methods for Managing Parkinson’s Disease. In *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*; Holzinger, A., Ed.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2016; pp 209–220. https://doi.org/10.1007/978-3-319-50478-0\_10.](https://www.zotero.org/google-docs/?nxPdeA)

[(13) YOON, S.; TAHA, B.; BAKKEN, S. Using a Data Mining Approach to Discover Behavior Correlates of Chronic Disease: A Case Study of Depression. *Stud. Health Technol. Inform.* **2014**, *201*, 71–78.](https://www.zotero.org/google-docs/?nxPdeA)

[(14) Longstreth, W. T.; Diehr, P.; Manolio, T. A.; Beauchamp, N. J.; Jungreis, C. A.; Lefkowitz, D. Cluster Analysis and Patterns of Findings on Cranial Magnetic Resonance Imaging of the Elderly: The Cardiovascular Health Study. *Arch. Neurol.* **2001**, *58* (4), 635–640. https://doi.org/10.1001/archneur.58.4.635.](https://www.zotero.org/google-docs/?nxPdeA)

[(15) Guthrie, E.; Creed, F.; Fernandes, L.; Ratcliffe, J.; Jagt, J. V. der; Martin, J.; Howlett, S.; Read, N.; Barlow, J.; Thompson, D.; Tomenson, B. Cluster Analysis of Symptoms and Health Seeking Behaviour Differentiates Subgroups of Patients with Severe Irritable Bowel Syndrome. *Gut* **2003**, *52* (11), 1616–1622. https://doi.org/10.1136/gut.52.11.1616.](https://www.zotero.org/google-docs/?nxPdeA)

[(16) Fitzpatrick, A. M.; Teague, W. G.; Meyers, D. A.; Peters, S. P.; Li, X.; Li, H.; Wenzel, S. E.; Aujla, S.; Castro, M.; Bacharier, L. B.; Gaston, B. M.; Bleecker, E. R.; Moore, W. C. Heterogeneity of Severe Asthma in Childhood: Confirmation by Cluster Analysis of Children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. *J. Allergy Clin. Immunol.* **2011**, *127* (2), 382-389.e13. https://doi.org/10.1016/j.jaci.2010.11.015.](https://www.zotero.org/google-docs/?nxPdeA)

[(17) Polat, H.; Danaei Mehr, H.; Cetin, A. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. *J. Med. Syst.* **2017**, *41* (4), 55. https://doi.org/10.1007/s10916-017-0703-x.](https://www.zotero.org/google-docs/?nxPdeA)

[(18) Salekin, A.; Stankovic, J. Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*; 2016; pp 262–270. https://doi.org/10.1109/ICHI.2016.36.](https://www.zotero.org/google-docs/?nxPdeA)

[(19) Qin, J.; Chen, L.; Liu, Y.; Liu, C.; Feng, C.; Chen, B. A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. *IEEE Access* **2020**, *8*, 20991–21002. https://doi.org/10.1109/ACCESS.2019.2963053.](https://www.zotero.org/google-docs/?nxPdeA)

[(20) Rady, E.-H. A.; Anwar, A. S. Prediction of Kidney Disease Stages Using Data Mining Algorithms. *Inform. Med. Unlocked* **2019**, *15*, 100178. https://doi.org/10.1016/j.imu.2019.100178.](https://www.zotero.org/google-docs/?nxPdeA)

[(21) UCI Machine Learning Repository: Chronic\_Kidney\_Disease Data Set http://archive.ics.uci.edu/ml/datasets/Chronic\_Kidney\_Disease (accessed Jun 22, 2020).](https://www.zotero.org/google-docs/?nxPdeA)

[(22) Han, J.; Pei, J.; Kamber, M. *Data Mining: Concepts and Techniques*; Elsevier, 2011.](https://www.zotero.org/google-docs/?nxPdeA)

[(23) Dash, M.; Liu, H. *IW’ELU ENT DATA ANALYSIS Www.Elsevier.ComAocate/Ida Feature Selection for Classification*; 1997.](https://www.zotero.org/google-docs/?nxPdeA)

[(24) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.](https://www.zotero.org/google-docs/?nxPdeA)

[(25) *Artificial Neural Network Modelling*; Shanmuganathan, S., Samarasinghe, S., Eds.; Studies in Computational Intelligence; Springer International Publishing: Cham, 2016; Vol. 628. https://doi.org/10.1007/978-3-319-28495-8.](https://www.zotero.org/google-docs/?nxPdeA)

[(26) Wolpert, D. H. Stacked Generalization. *Neural Netw.* **1992**, *5* (2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.](https://www.zotero.org/google-docs/?nxPdeA)

[(27) Wu, J. Cluster Analysis and K-Means Clustering: An Introduction. In *Advances in K-means Clustering: A Data Mining Thinking*; Wu, J., Ed.; Springer Theses; Springer: Berlin, Heidelberg, 2012; pp 1–16. https://doi.org/10.1007/978-3-642-29807-3\_1.](https://www.zotero.org/google-docs/?nxPdeA)

[(28) Pagnuco, I. A.; Pastore, J. I.; Abras, G.; Brun, M.; Ballarin, V. L. Analysis of Genetic Association Using Hierarchical Clustering and Cluster Validation Indices. *Genomics* **2017**, *109* (5), 438–445. https://doi.org/10.1016/j.ygeno.2017.06.009.](https://www.zotero.org/google-docs/?nxPdeA)

[(29) Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27* (8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.](https://www.zotero.org/google-docs/?nxPdeA)