International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (Jan-2022)

https://dx.doi.org/10.12785/ijcds/110113

A Predictive Analysis of Chronic Kidney Disease by Exploring Important Features

Mafizur Rahman[1], Linta Islam[2], Masud Rana[3], Malika Zannat Tazim[4], Jannatul Ferdous Sorna[4] and Syada Tasmia Alvi[5]

[1,3,4]*Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh*
[2]*Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh*
[5]*Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh*

Received 15 Mar. 2021, Revised 9 Aug. 2021, Accepted 13 Nov. 2021, Published 9 Jan. 2022

Abstract: Chronic Kidney Disease is an incurable disease which causes damages to the functions of a kidney gradually. Only proper treatment can prevent the disease from getting worse. Because of proper knowledge about kidney disorders, people had to suffer from this deadly disease. Thus, in this paper, we analyzed certain key features and noticed several interesting relationships with the disease by considering the actual perception of people. We also predict kidney disease by employing various machine learning algorithms including Logistic Regression, Naive Bayes, SVM and KNN. By applying PCA, we observe that there is an improvement in the accuracy for predicting the disease. SVM outperforms other algorithms with 98% accuracy in predicting chronic kidney disease. In future, we will try to find some significant hypothesis that helps us to prevent the disease better.

Keywords: Kidney Disease, PCA, Null Hypothesis, Accuracy, Correlation, Features

1. Introduction and Overview

Kidney failure is becoming a significant issue in this present world. Nowadays, without feeling any symptoms or complications, kidney function may decrease up to 90 percent [1]. Chronic Kidney Disease (CKD) additionally known as Chronic Renal Disease (CRD) that acts as a silent killer. Several causes, even including scarcity of energy, drowsiness, exhaustion, pruritus, pain may have CKD affected patients [2]. Furthermore, smoking, diabetes, abnormally high bp, cardiac disease, obesity, renal disorder family history, consumption of alcohol, age, race, male sex, and drug addiction may increase the prospect of kidney disease [3]. Moreover, the dysfunction of the renal affects the entire human body even it may be the cause of extreme illness and death. According to the kidney international report, among the top five serious diseases, CKD is the top concern that can cause the death of a human being [4]. Perhaps the absence of a commonly used instrument for prediction seems to have become a bigger concern for the growing number of CKD patients [5].

Therefore, it is necessary to detect this disease in the early stages. Thus, researchers need to reveal the crucial information regarding this disease to take preventive action before the disease reaches the severe stage. In these circumstances, data mining and machine learning applications are playing a significant role in medical research by identifying the proper information of any illnesses [5]. Several factors lie behind this disease and it is often hard to find out the significant factors that influence this disease [6]. Thus, proper analysis of the important features of this disease can carry out considerable knowledge regarding this disease in medical research. Hence, the main motive of this paper was to predict kidney disease by exploring its crucial factors. The author deeply analyzed the various relationship between kidney disease factors by generating few hypotheses, therefore, it will help the medical practitioner to take preventive action easily.

The research objectives are:

- Predict kidney disease by using classification techniques with dimensionality reduction.
- To analyze the key features of kidney disease from the dataset.
- To explore the relationship between the important features with kidney disease by generating hypotheses.

This research has been designed to illustrate a predictive analysis of chronic kidney disease by revealing the impactful features of this disease. To analyze the features, we surveyed those who came to Kidney Foundation Hospital and

E-mail address: rmafizur10@gmail.com, lislamcsejnu@gmail.com, masudewu2016@gmail.com, zannattazim1996@gmail.com, sornaewu@gmail.com, syada.cse@diu.edu.bd

Research Institute Bangladesh for treatment. This institute started its journey in 2003 and at present, it is one of the largest treatment centers for kidney patients in Bangladesh.

2. Related Work

Day by day, researchers are trying to discover new information regarding various medical disease. Several machine learning applications has been utilized to analyze and predict the behavior of kidney diseases [7]. Among these algorithms, Using the WEKA tool, they showed that Decision Tree gives the best result, and there is a meaningful relationship between accuracy and the varying attribute. The authors also suggest that to detect kidney diseases raking algorithm should be used [7]. Rady et al. [8] conducted which algorithm gives the best result to detect CKD of several stages of the patient. The authors also concluded that the Probabilistic Neural Networks algorithm (PNN) algorithm predict at best accuracy at every critical stage of the patient to predict CKD. To detect CKD at an early stage, an empirical result [9] showed that ANN performed better compared to SVM to predict the early stage of CKD.

There was a motive to find a solution to the prevalence from the normal stage to the critical stage of CKD. To find the solution, the authors [10] applied the Naive Bayes algorithm with OneR. The researchers also suggested using their action rules at the respective stage of CKD. Multivariate statistical techniques were applied by Charleonnan et al. [11] because several danger factors were linked to make a good prediction of CKD. Finally, the authors concluded that Clustering Heatmap gives a predictive model of healthcare administration that must reduce the risk of CKD. Zeynu and Patil [12] proposed two models Feature selection and ensemble model to predict the CKD. They applied the Wrapper Subset and Ranker Search Engine to find the gain of the attributes. Apart from this, the authors also applied different machine learning algorithms and built a second method by combing the five heterogeneous classifiers based on a voting algorithm.

Papers [11], [10], [13], [8], [7] investigated kidney disease by utilizing several machine learning and deep neural network [9] along with feature selection models [12]. These papers addressed numerous heterogeneous classifiers and probabilistic approaches to analyze the features. The main limitation was to make a better prediction, the authors did not consider all of the 25 features from the dataset. Therefore, this study is designed by considering all of the features from the dataset and also, explored important features with a new scheme.

3. Methodology

In this study, we utilized several classification algorithms to predict kidney disease. Then, we explored the features for which we achieved high accuracy. To explore those features, we took a questionnaire survey to get the actual perception about those features for kidney disease. Figure 1 shows the system architecture of our proposed work.

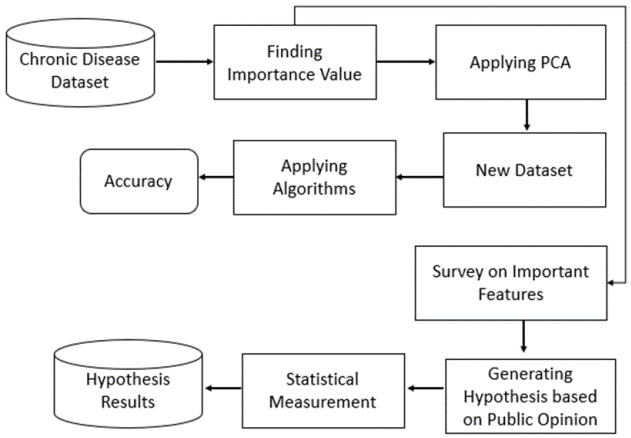

Figure 1. Workflow Diagram

To proceed with this task, we have collected the kidney disease dataset. By extracting the value for each feature, PCA has been applied for reducing the dimension of the dataset. The classification algorithms are applied to the newly formed dataset to get the maximum accuracy. Based on the importance value, the main features have been extracted and a survey has been conducted to explore those features towards people perception who are related to the disease. To analyze features more deeply, we designed several hypotheses and measured the hypotheses value statistically.

We collected the dataset from the UCI machine learning repository for kidney disease classification [14]. The dataset contains 25 attributes with 400 instances. We removed the irrelevant punctuations, null values, and other ambiguous data to classify with high accuracy. There were two class labels 'CKD' and 'NotCKD' where CKD implies the person has kidney disease whereas NotCKD implies the person has no kidney disease. Below is the list of algorithms that we utilized for our classification purpose.

i) Naive Bayes:

A Naive-Bayes classifier is a basic probabilistic system based on the Bayes law including a strict assumption of independence. These classifiers are thoroughly scalable, requiring a set of linear parameters for the number of variables in a learning problem (features/predictors). In most other complex real-world scenarios, Naive-Bayes classifiers have performed very well, notwithstanding their simplistic nature and seemingly oversimplified assumptions. A review of such a probabilistic (Bayesian) classification problem shows that the seemingly unplausible effectiveness of naive Bayes classifiers has sound theoretical explanations [15]. Basically, Naive-Bayes is a kind of a model of conditional probability.

ii) K-Nearest Neighbors (KNN):

The K-Nearest-Neighbors (KNN) algorithm is a method of supervised Machine-Learning (ML) algorithm that can be used for both regression and classification predictive issues. The K- Nearest-Neighbors (KNN) algorithm uses similarities of features for predicting new DP (data point) values, which also suggests that value would be given to the new data point depending on how exactly it fits the points in the training set. The appropriate value of k depending on data; typically, higher values of k decrease the influence of noise [16] on the classification but making the boundaries fewer distinct between groups.

iii) Support Vector Machine (SVM):

SVM is basically a very powerful flexible supervised Machine- Learning technique. Though its uses for classification issues and regression issues, it is popular for predicting classification issues. Compared to different existing Machine Learning algorithms, SVM has its own special method of implementation. Lately, due to its capability to manipulate multiple categorical and continuous variables, it is incredibly popular. Moreover, in multi-dimensional space, an SVM network is primarily a visualization of various groups in a hyperplane [17]. To search for a maximal marginal hyperplane (MMH), the purpose of SVM is to split datasets into different classes.

iv) Logistic Regression:

Logistic regression is a statistical model that uses a logistic feature to simulate a binary dependent variable in its simplest form. It is a widely used method for classification. This is a multiple regression variant where the expected result is binary instead of quantitative [18]. In different fields of medical research, logistic regression is vastly used. Using logistic regression, several medical scales that are used to determine a patient's severity, have been developed [19]. In order to determine the probability of developing a given condition like kidney or coronary heart disease, logistic regression can also be used depending on the patient's observed characteristics such as patient's age, sex, body mass index, results of various blood tests, etc. [20].

Principal Component Analysis: Principal component analysis (PCA) is a method to lessen the dimensionality [21] of the dataset for creating predictive paradigms by exploratory analysis [22]. This process can transform data into a different coordinate by conserving the variations in the data. There will be a lower amount of information loss because of the higher interpretability rate. For example, if our 3-dimensional dataset contains 3 types of variables. Hence, there will be three eigenvectors that include three eigenvalues and those eigenvalues measure the variance that is carried by each principal component. The 1st principal component always refers to the projections which have the highest variance in the space direction. The 2nd component also maximizes the variation in the orthogonal directions of the 1st component.

We have employed PCA techniques in our study to reduce the data dimensionality for better prediction of kidney disease. We have utilized a dataset that contains 25 attributes. Hence, when we have calculated the importance value for each of them, we find a low importance value for a few features that indicate those features have low relation to our output. Hence, we have consolidated those features into two principal components that reduced the dataset dimension. As the dataset is small, therefore, feature selection may lead to an overfitting problem. To avoid this problem, we choose the features based on their importance value and utilized 10-fold cross-validation for testing and training purposes.

Feature importance: Feature importance is a score that indicates the relative weight of every feature for making a prediction. To get the importance value of each feature, we utilized the ExtraTreesClassifier ensemble method. By following classical top-down procedure this classifier can create an ensemble of the unpruned decision tree by splitting nodes. The normalizing reduction in the Gini index helps in the decision of splitting features and provides the Gini Importance value for each feature.

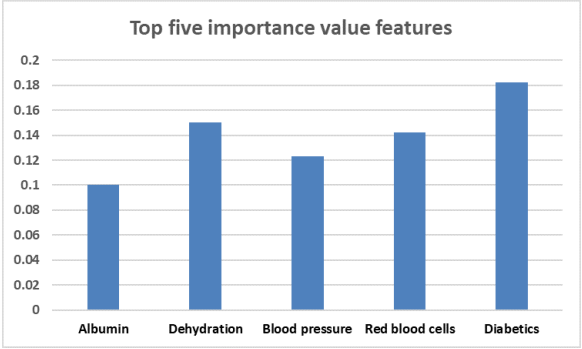

Figure 2. Highest importance value features

Figure 2 shows the most impactful features are albumin, dehydration, blood pressure, red blood cells, and diabetics. The importance values of these features are 0.10, 0.15, 0.125, 0.142 and 0.182. We found the importance value for seven features (bacteria, potassium, coronary artery disease, age, anemia, blood glucose random, and sugar) are less than 0.015 which is very low.

These seven are the least impactful features in the dataset. Thus, we made two new components where component 01 includes bacteria, potassium, coronary artery disease, and component 02 include age, anemia, blood glucose random, and sugar that shows in figure 3. After that, we merged the data into the dataset and the dimension of the dataset got reduced. The new dataset contains 19 features and we proceed to the next stage to compute the accuracy for prediction.

Data Collection Strategy: To explore the important features of kidney disease, we have picked the questionnaire

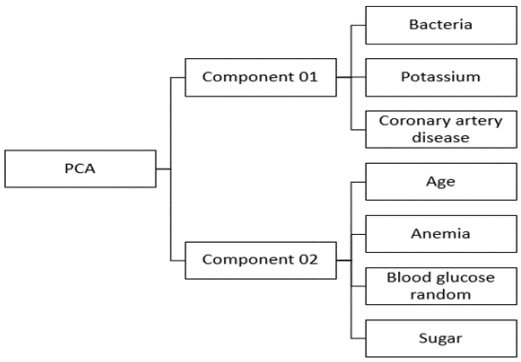

Figure 3. Component Category through PCA

approach as our data collection strategy. This method is utilized to carry out something by inquiring people about a subject or topics [23]. This method helps to manage data in a short time through a survey form and it is also cost-effective for us for convenient research [24]. To proceed with our work, we have visited the Kidney Foundation Hospital and Research Institute Bangladesh and later we have surveyed only particular person who have come there for treatment. We have given the survey form to them to give their personal opinion about kidney disease features. As we did a non-random sampling therefore, by employing a pragmatic way, small sample can generate a good result. Due to the pandemic situation, we failed to gather a large amount of data. Therefore, we have collected 170 data and the total amount of questions was 10.

Ethics: To collect personal data we have visited the Kidney Foundation Hospital and Research Institute Bangladesh which is one of the largest dialysis and transplant center in Bangladesh. It is also a cost-effective hospital that provides services with an affordable budget. Since we have collected personal data, hence, we have provided an informed consent form to our participants before collecting data. We have given sufficient information about this task. We have also assured to keep their secrecy and their data will not be handover to third parties for further use.

Null Hypothesis: In the area of statistics, the statement that evaluates there is no association within two measured aspects is known as the null hypothesis [25]. The purpose of this null hypothesis is to find out the comparative explanation of the aspects. To test the hypothesis, we consider the null hypothesis is true for a distinct data set. The null hypothesis can be accepted or rejected [26]. It depends on the statistical significance rate known as the P-value. The range of this p-value is 0 to 1. If the P-value is very small then it concludes that the evidence is too strong towards the null hypothesis [27]. Therefore, the null hypothesis should be rejected.

We have calculated the importance value for each of the features for identifying kidney disease from our utilized dataset. Therefore, we tried to test the relationship between those features with kidney disease in terms of people's perceptions. We have designed some questions adopting those features (albumin, dehydration, blood pressure, red blood cells, and diabetics) and collected the data by doing a survey. After that, we have generated a few null hypotheses as well as the alternative hypotheses to bring out the association between the features which has larger importance value towards kidney disease. To evaluate the relationship between kidney disease and it's features in terms of people's perceptions, we have designed six hypotheses with their alternative (Alt.) hypotheses through a case study. We took the best features which have greater importance value towards this disease. Then, we did a quantitative analysis to reveal how people think about those important features who have affected any kind of kidney disorder and whether those features are valuable or not in real life perspective. Hypotheses are:
1) H0: Affected people think blood pressure has no relation to kidney disease.
Alt. H0: Affected people think blood pressure has related to kidney disease.
2) H1: Affected people are not aware of blood pressure.
Alt. H1: Affected people are aware of blood pressure.
3) H2: Affected people think diabetes has no connection to kidney disease.
Alt. H2: Affected people think diabetes has a connection to kidney disease.
4) H3: Affected people didn't feel dehydrated for this disease.
Alt. H3: Affected people felt dehydrated for this disease.
5) H4: Affected people think albumin has no relation to kidney disease.
Alt. H4: Affected people think albumin is related to kidney disease.
6) H5: Affected people don't know about the function of red blood cells.
Alt. H5: Affected people know about the function of red blood cells.

1) Chi-Square Analysis

A Chi-square analysis has conducted to discover the connection between two categorical variables. This test gives us whether the relationship between the two variables is significant or not. To complete this analysis, we followed few steps described below.

- **Step 1:** We have state seven hypotheses as well as their alternative hypotheses and also designed the contingency table for each hypothesis.
- **Step 2:** We have formulated the analysis plan by utilizing 0.05 as our significance levels. Yet, most of the researchers employ 0.01 or 0.10 as significance levels.
- **Step 3:** To analyze our data, we have calculated the degree of freedom. After that, we have interpreted

the expected frequencies, and finally, the test statistics value.

1) Degree of Freedom (DF): The DF can be calculated as follow:

$$DF = (r-1)*(c-1)$$

Here, r and c imply the number of levels for first and second categorical variable.

2) Expected Frequency: We have calculated this frequency separately for every level of the categorical variables according to the following equation:

$$E_{r,c} = \frac{(n_r * n_c)}{n}$$

Here, Total number of rows = n_r
Total number of column = n_c
Entire sample size = n

3) Test Statistic: We have calculated the test statistic for chi-square random variable X^2 by the following formula:

$$X^2 = \sum \frac{(O_{r,c} - E_{r,c})^2}{E_{rc}}$$

Where, $O_{r,c}$ is the perceived frequency for both variables and $E_{r,c}$ is the expected frequency for both variables.

- **Step 4:** At last, we have calculated the P-value that shows the probability of perceiving statistical samples as far as the test statistics. Finally, we examined the P-value with the level of significance to interpret our hypotheses.

The collected medical data is used only for hypothesis purposes for revealing a few relations among the top features and the topmost features are selected from the CKD dataset by employing machine learning algorithms. As the data is collected from actual patients, hence, the hypotheses results will provide them with a better conception regarding their disease features.

4. RESULTS AND FINDINGS

Correlation of the Features: Correlation is a mathematical concept in common use that refers to how similar 2 variables are to having a linear relationship with each other. Mostly as the preprocessing phase for machine learning, feature selection is efficient in reducing the dimension of the feature, eliminating unnecessary data, increasing the accuracy of learning, and improving the understandability of performance. A feature or function is useful if it is associated with or the class' predictive; otherwise, it becomes meaningless. A fast filter method that can classify relevant characteristics as well as similarity among relevant characteristics without the study [28]. Hence, for measuring the similarity between the features, we have generated a correlation matrix of the features.

Figure 4 displays the correlation matrix plot for kidney disease features where yellow color indicates a strong relationship and Figure 5 presents the values of the matrix for the first 14 features as an example.

TABLE I. Results Before Applying PCA (25 Attributes)

	Precision	Recall	Accuracy(%)
Logistic Regression	0.874	0.851	84
Naive Bayes	0.860	0.831	85
SVM	0.830	0.861	84
KNN	0.794	0.812	79

TABLE II. Results After Applying PCA (19 Attributes)

	Precision	Recall	Accuracy (%)
Logistic Regression	0.961	0.970	96
Naive Bayes	0.956	0.972	95
SVM	0.970	0.981	98
KNN	0.948	0.922	93

Figure 6 displays the conditional relationship for some features in a multiplot grid. The grid plot has been represented with the hue parameter for better visualization.

Kidney Disease Prediction: We employed four classification algorithms to predict kidney disease. Table I implies that the highest accuracy 85% achieved from the Naive Bayes algorithm before dimensionality reduction. After that, when we reduced the dataset dimension, the prediction accuracy increased for all the models.

Table II shows that we achieved 98% accuracy using the SVM technique that outperformed the rest of the models.

Highly imbalanced data into positive and negative classes can generate high accuracy. As it will observe based on the majority class and predict the output which could be wrong with high accuracy. Therefore, after getting the final model, we also calculated the ROC AUC which is 99.2%. Figure 7 shows the roc curve area for our most competent model for the threshold value 0 to 1.

Comparison: In 2019, the authors [13] proposed an intelligence system by employing an Ant Colony based Optimization (D-ACO) framework to predict chronic kidney disease with a density-based feature selection model. To increase the accuracy, the authors used 16 features and achieved the highest accuracy 95% whereas we achieved 98% accuracy using 19 attributes. To maximize the accuracy, we have consolidated the lower features into two components without dropping them from the dataset. Thus, this study also outperformed the previous work.

Hypotheses of the Study:

H0: Affected people think blood pressure has no relation

	Age	Bp	Sg	Al	Su	Rbc	Pc	Pcc	Ba	Bgr	...	Hemo	Pcv	Wbcc	Rbcc
Age	1.000000	0.148001	-0.159071	0.114845	0.207851	-0.062675	-0.138997	0.156844	0.041881	0.214408	...	-0.175371	-0.211807	0.100061	-0.201100
Bp	0.148001	1.000000	-0.164423	0.146654	0.200781	-0.197868	-0.170790	0.057634	0.110767	0.149100	...	-0.279522	-0.292724	0.026067	-0.220804
Sg	-0.159071	-0.164423	1.000000	-0.460148	-0.277489	0.329399	0.356355	-0.305094	-0.230662	-0.317896	...	0.492103	0.501080	-0.206884	0.443437
Al	0.114845	0.146654	-0.460148	1.000000	0.268069	-0.381881	-0.547316	0.397016	0.365788	0.326419	...	-0.549871	-0.527369	0.200402	-0.454735
Su	0.207851	0.200781	-0.277489	0.268069	1.000000	-0.118402	-0.188849	0.149968	0.106567	0.639185	...	-0.191498	-0.202574	0.153276	-0.182173
Rbc	-0.062675	-0.197868	0.329399	-0.381881	-0.118402	1.000000	0.360892	-0.090962	-0.161818	-0.196875	...	0.371731	0.353885	0.004913	0.265451
Pc	-0.138997	-0.170790	0.356355	-0.547316	-0.188849	0.360892	1.000000	-0.502572	-0.320167	-0.275167	...	0.458642	0.454645	-0.111617	0.413157
Pcc	0.156844	0.057634	-0.305094	0.397016	0.149968	-0.090962	-0.502572	1.000000	0.274493	0.195641	...	-0.273431	-0.290826	0.162689	-0.263150
Ba	0.041881	0.110767	-0.230662	0.365788	0.106567	-0.161818	-0.320167	0.274493	1.000000	0.084449	...	-0.203163	-0.187253	0.102943	-0.188703
Bgr	0.214408	0.149100	-0.317896	0.326419	0.639185	-0.196875	-0.275167	0.195641	0.084449	1.000000	...	-0.269134	-0.267595	0.121370	-0.222459
Bu	0.187541	0.183969	-0.249371	0.405921	0.152178	-0.243188	-0.383153	0.182395	0.156929	0.127491	...	-0.540635	-0.525994	0.041510	-0.465892
Sc	0.127324	0.144359	-0.176141	0.230678	0.130115	-0.164996	-0.205440	0.048394	0.049689	0.082241	...	-0.342053	-0.341873	-0.005418	-0.323056
Sod	-0.085948	-0.103217	0.217456	-0.271418	-0.073705	0.181556	0.200728	-0.140562	-0.080550	-0.154388	...	0.333604	0.346820	0.006334	0.316883
Pot	0.050153	0.066671	-0.063450	0.114470	0.181048	0.032226	-0.155536	-0.006580	-0.002934	0.056757	...	-0.100612	-0.123305	-0.074057	-0.120418
Hemo	-0.175371	-0.279522	0.492103	-0.549871	-0.191498	0.371731	0.458642	-0.273431	-0.203163	-0.269134	...	1.000000	0.854994	-0.153810	0.681864
Pcv	-0.211807	-0.292724	0.501080	-0.527369	-0.202574	0.353885	0.454645	-0.290826	-0.187253	-0.267595	...	0.854994	1.000000	-0.183397	0.703360
Wbcc	0.100061	0.026067	-0.206884	0.200402	0.153276	0.004913	-0.111617	0.162689	0.102943	0.121370	...	-0.153810	-0.183397	1.000000	-0.151380

Figure 4. Matrix of Correlation Between Column

TABLE III. Blood pressure towards kidney disease

Affected Person	Blood Pressure	
	YES	**NO**
Male	11	87
Female	9	63

TABLE IV. BP awareness towards kidney disease

Affected Person	BP Awareness		
	YES	**NO**	**SOMETIMES**
Male	47	25	26
Female	42	9	21

TABLE V. Diabetes relation to kidney disease

Affected Person	Diabetes	
	YES	**NO**
Male	78	20
Female	60	12

TABLE VI. Dehydration impact towards kidney disease

Affected Person	Dehydrated	
	YES	**NO**
Male	92	06
Female	64	08

to kidney disease.
Alt. H0: Affected people think blood pressure has related to kidney disease.

Table III shows among all respondents, a substantial portion of the affected male and female people think blood pressure has no relation to kidney disease whereas only a few think it has a relation to kidney disease.

H1: Affected people are not aware of blood pressure.
Alt. H1: Affected people are aware of blood pressure.

This study reveals that once a person got affected with kidney disease then they become aware of their blood pressure. Table IV refers among 98 affected males, 73(Yes=47, Sometimes=26) males aware of it and among 72 affected female participants, 63(Yes=42, Sometimes=21) females are also conscious of it.

H2: Affected people think diabetes has no connection to kidney disease.
Alt. H2: Affected people think diabetes has a connection to kidney disease

Table V comprises 78 male and 60 female participants who are affected, think diabetes has a connection to kidney disease, and the rest of the participants believe it has no relation to diabetes.

H3: Affected people didn't feel dehydrated for this disease.
Alt. H3: Affected people felt dehydrated for this disease.

The respondents were asked about whether they felt dehydration when they were suffering from kidney disease. Table VI shows 92 male and 64 female participants answered that they felt dehydration.

H4: Affected people think albumin has no relation to kidney disease.
Alt. H4: Affected people think albumin is related to kidney disease.

From Table VII, 42 males and 37 females answered that albumin has relation with kidney disease whereas 19(13+6) persons don't think so and the rest of the participants sometimes believe that albumin has a relation to this disease.

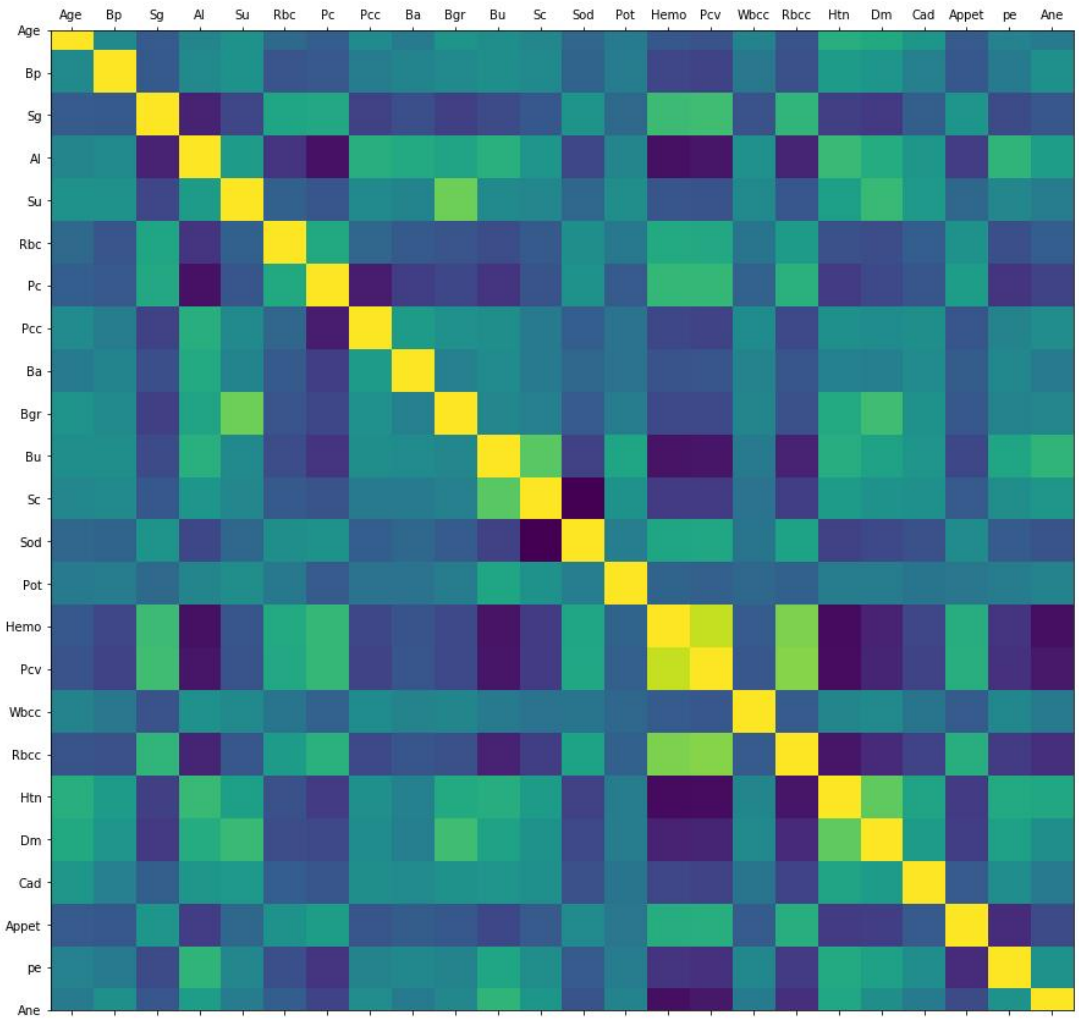

Figure 5. Correlation Matrix

TABLE VII. People's concern about Albumin

Affected Person	Taking Protein		
	YES	**NO**	**SOMETIMES**
Male	42	13	43
Female	37	6	29

TABLE VIII. Knowledge about red blood cells function

Affected Person	RBC knowledge		
	YES	**NO**	**Little**
Male	58	7	33
Female	47	7	18

H5: Affected people don't know about the function of red blood cells.
Alt. H5: Affected people know about the function of red blood cells.

Amongst 170 participants, Table VIII confers 91(Yes=58, Little=33) male and 65(Yes=47, Little=18) female responses that they know the functions of red blood cells. Only a few people do not know about it.

Results of the Chi-square analysis: From chi-square analysis for significance level 0.05, we found the hypotheses are statistically not significant. Table IX shows the output of hypotheses after the chi-square test. It shows for all the hypotheses the P values are greater than the significance value.

Therefore, the null hypotheses will not reject in this circumstance. This result confers that the relationship among the variables which is observed in the sample is expected to befall by chance. There could be few limitations as the

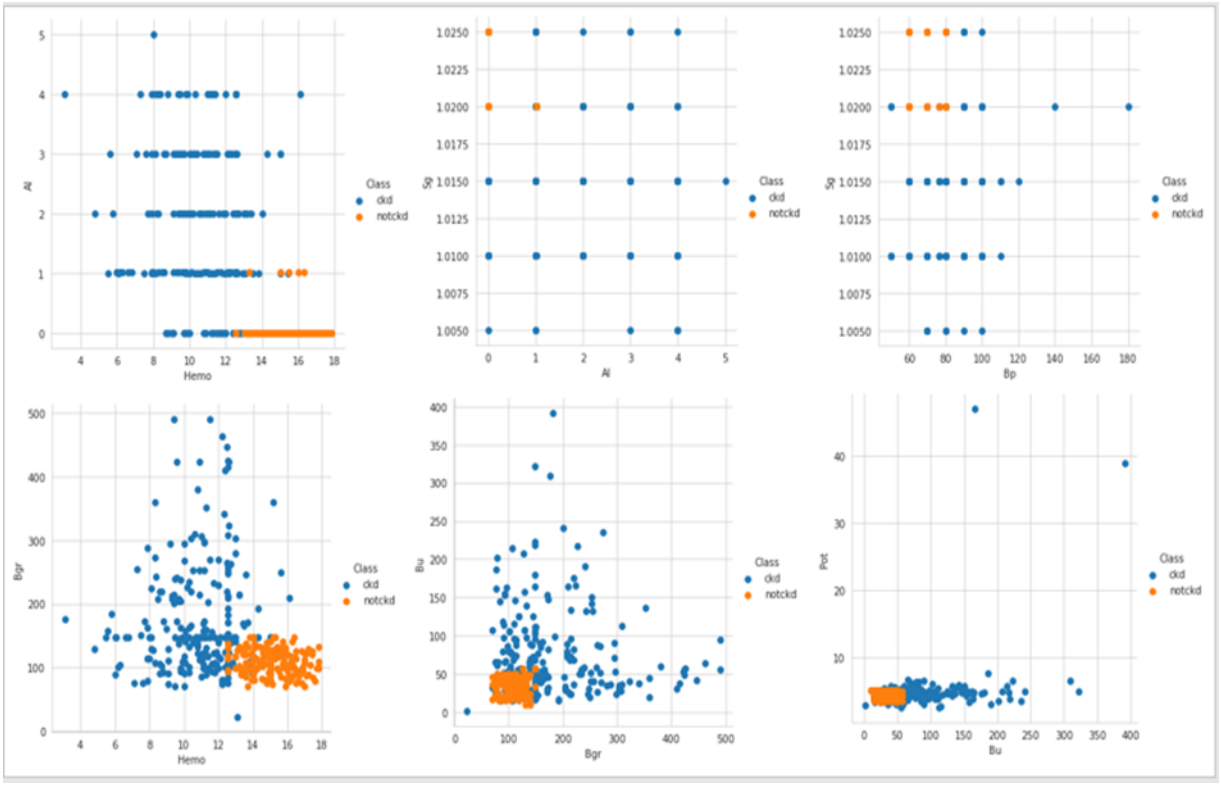

Figure 6. Conditional relationship between the features

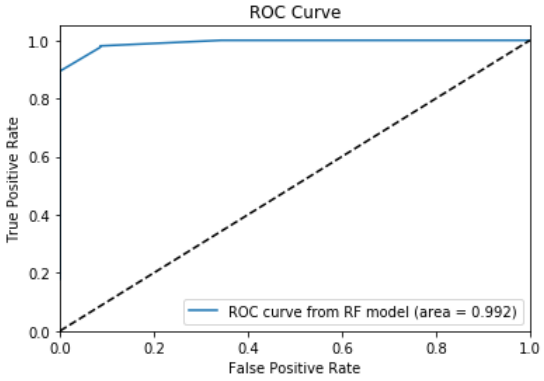

Figure 7. Roc Curve for the Best Model

TABLE IX. Chi-square analysis output

	Chi Value	**P Value**	**S/NS**
Hypothesis 01	0.065	0.79	Not Significant
Hypothesis 02	4.47	0.11	Not Significant
Hypothesis 03	0.38	0.53	Not Significant
Hypothesis 04	1.36	0.24	Not Significant
Hypothesis 05	1.68	0.43	Not Significant
Hypothesis 06	1.62	0.44	Not Significant

sample size is not so large and the variability of the data is high in random sampling, thus it will have an impact on the results. The author worked on small scall data for finding the relation between various features which is statistically not significant for all the cases. It means there is a lack of evidence for these cases and this is because of a small sample. The results can be improved by taking a large amount of sampling data that will produce more reliable consequences.

5. Conclusions

In medical research, data mining and machine learning have shown great effectiveness through predictive analysis. As it can extract important information from large data, thus, it helps medical researchers to find out the proper solution for a complex disease. This study focused on the classification of kidney disease to predict whether a person has this disease or not. The authors achieved 98% accuracy by utilizing the SVM technique. As we worked with the important feature by consolidating low important

features through the PCA technique the algorithms work well and also discard the overfitting problem through cross-validation. The importance values of all 25 features were extracted with the ExtraTreesClassifier ensemble approach and finally made a dataset with 19 features. This study also reveals the relation of kidney disease features towards the disease more deeply and how these features impact kidney failure through people's perceptions. In the future, we will try new models to reveal more information regarding this disease with greater accuracy.

References

[1] J. Chen, R. P. Wildman, D. Gu, J. W. Kusek, M. Spruill, K. Reynolds, D. Liu, L. Hamm, P. K. Whelton, and J. He, "Prevalence of decreased kidney function in chinese adults aged 35 to 74 years," *Kidney international*, vol. 68, no. 6, pp. 2837–2845, 2005.

[2] S. J. Lee and J. Jeon, "Relationship between symptom clusters and quality of life in patients at stages 2 to 4 chronic kidney disease in korea," *Applied Nursing Research*, vol. 28, no. 4, pp. e13–e19, 2015.

[3] J. Low, G. Smith, A. Burns, and L. Jones, "The impact of end-stage kidney disease (eskd) on close persons: a literature review," *NDT plus*, vol. 1, no. 2, pp. 67–79, 2008.

[4] J. Radhakrishnan and S. Mohan, "Ki reports and world kidney day," *Kidney international reports*, vol. 2, no. 2, pp. 125–126, 2017.

[5] M. Fatima, M. Pasha *et al.*, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 01, p. 1, 2017.

[6] M. A. Bruce, B. M. Beech, M. Sims, T. N. Brown, S. B. Wyatt, H. A. Taylor, D. R. Williams, and E. Crook, "Social environmental stressors, psychological factors, and kidney disease," *Journal of Investigative Medicine*, vol. 57, no. 4, pp. 583–589, 2009.

[7] N. Tazin, S. A. Sabab, and M. T. Chowdhury, "Diagnosis of chronic kidney disease using effective classification and feature selection technique," in *2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec)*. IEEE, 2016, pp. 1–6.

[8] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, p. 100178, 2019.

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

[10] U. N. Dulhare and M. Ayesha, "Extraction of action rules for chronic kidney disease using naïve bayes classifier," in *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2016, pp. 1–5.

[11] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *2016 management and innovation technology international conference (MITicon)*. IEEE, 2016, pp. MIT–80.

[12] S. Zeynu and S. Patil, "Prediction of chronic kidney disease using data mining feature selection and ensemble method," *International Journal of Data Mining in Genomics & Proteomics*, vol. 9, no. 1, pp. 1–9, 2018.

[13] M. Elhoseny, K. Shankar, and J. Uthayakumar, "Intelligent diagnostic prediction and classification system for chronic kidney disease," *Scientific reports*, vol. 9, no. 1, pp. 1–14, 2019.

[14] A. Frank, "Uci machine learning repository," *http://archive. ics. uci. edu/ml*, 2010.

[15] H. Zhang, "Exploring conditions for the optimality of naive bayes," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 02, pp. 183–198, 2005.

[16] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Cluster analysis 5th ed," 2011.

[17] D. Caragea, D. Cook, H. Wickham, and V. Honavar, "Visual methods for examining svm classifiers," in *Visual Data Mining*. Springer, 2008, pp. 136–153.

[18] K. Kempf-Leonard, "Encyclopedia of social measurement," 2004.

[19] M. Kologlu, D. Elker, H. Altun, and I. Sayek, "Validation of mpi and pia ii in two different groups of patients with secondary peritonitis." *Hepato-gastroenterology*, vol. 48, no. 37, pp. 147–151, 2001.

[20] D. A. Freedman, *Statistical models: theory and practice*. cambridge university press, 2009.

[21] M. Rahman, M. M. Zahin, and L. Islam, "Effective prediction on heart disease: Anticipating heart disease using data mining techniques," in *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, 2019, pp. 536–541.

[22] D. Chernyshov, I. Dovgaliuk, V. Dyadkin, and W. van Beek, "Principal component analysis (pca) for powder diffraction data: Towards unblinded applications," *Crystals*, vol. 10, no. 7, p. 581, 2020.

[23] M. R. Gecht, K. J. Connell, J. M. Sinacore, and T. R. Prohaska, "A survey of exercise beliefs and exercise habits among people with arthritis," *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, vol. 9, no. 2, pp. 82–88, 1996.

[24] R. Newby, J. Watson, and D. Woodliff, "Sme survey methodology: Response rates, data quality, and cost effectiveness," *Entrepreneurship Theory and Practice*, vol. 28, no. 2, pp. 163–172, 2003.

[25] R. L. Hagen, "In praise of the null hypothesis statistical test." 1997.

[26] R. J. Millikin, M. R. Shortreed, M. Scalf, and L. M. Smith, "A bayesian null interval hypothesis test controls false discovery rates and improves sensitivity in label-free quantitative proteomics," *Journal of proteome research*, vol. 19, no. 5, pp. 1975–1981, 2020.

[27] A. J. Harrison, S. A. McErlain-Naylor, E. J. Bradshaw, B. Dai, H. Nunome, G. T. Hughes, P. W. Kong, B. Vanwanseele, J. P. Vilas-Boas, and D. T. Fong, "Recommendations for statistical analysis involving null hypothesis significance testing," 2020.

[28] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.

Mafizur Rahman Mafizur Rahman received his BSc degree in computer science and engineering from East West University, Dhaka, Bangladesh. He worked as an undergraduate teaching assistant during his bachelor's program. His research interest lies in machine learning, data mining, and crowdsourcing.

Linta Islam Linta Islam received her B.Sc and M.Sc in Computer Science and Engineering, Jagannath University,Bangladesh. Now,she is working as a lecturer in the same department. She does research in Information Science, Computer Security and-Crowdsourcing. She has published 28 papers in various journals and conferences. In her supervision, several papers have also been published in IEEE conferences.

Masud Rana Masud Rana completed his BSc degree in CSE from East West University, Dhaka, Bangladesh. He worked on several machine learning models and published several conference papers. He received a merit scholarship award at EWU.

Malika Zannat Tazim Malika Zannat Tazim Completed her graduation in CSE from East West University, Bangladesh. She published several international conference papers. Her research interest lies in blockchain and machine learning.

Jannatul Ferdous Sorna Jannatul Ferdous Sorna graduated from East West University, Bangladesh. She worked as an undergraduate teaching assistant at EWU and also published few conference papers. Her research area lies in data mining and machine learning.

Syada Tasmia Alvi Syada Tasmia Alvi received her B.Sc in Computer Science and Engineering, Jagannath University, Bangladesh. Now she is working as a lecturer at Daffodil International University, Bangladesh. She does research in Computer Security, Crowdsourcing and Blockchain.