*Grenze International Journal of Engineering and Technology, Jan Issue*

# KNN Imputation with Correlation of Classifiers for Corroborate Prognosis of Chronic Kidney Disease (CKD)

T. Saroja[1] and Dr. Y. Kalpana[2]
[1]Research Scholar, VISTAS, Pallavaram, Assistant Professor, A.M.Jain College, Meenmabakkam, Chennai
Email: tsaroja11@gmail.com, saroja.t@amjaincollege.edu.in
[2]Research Supervisor, Professor Dept of IT, VISTAS, Pallavaram,Chennai
Email: kalpana.scs@velsuniv.ac.in

***Abstract*— Nowadays, the major health problem seen globally is Chronic Kidney Disease (CKD). It has high morbidity and mortality rate, and it induces other diseases. Unlike other diseases, this doesn't show any symptoms in the early stage, so patients often fail to notice the disease. Early detection of CKD enables patients to receive timely treatment to enhance the progression of this disease. The major challengence faced by physician is mostly patients fail to miss some medical history. Machine learning models can effectively help physicians achieve this goal due to their fast and accurate recognition performance. The datasets which contain medicial history of the patients contain irrelevant features that may negatively affect the disease diagnosis efficiently. Inorder to avoid these problems the datasets are imputed using various imputation methods. Then the imputed data are fed to classifiers to observe the performance of the model. The commonly used machine learning classifier methods are Logistic Regression (LOG), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NBs) and Feed Forward Neural Network (FFNN).**

**In this study K Nearest Neighbor (KNN) imputation was used to fill in the missing values, which selects several complete samples with the most similar measurements to process the missing data for each incomplete sample. Then the performance is observed by feeding the data in various classifiers and the results are compared against each type. The CKD dataset was obtained from the University of California Irvine (UCI) machine learning repository, which has a large number of missing values. The classifier and existing methods are evaluated using the metrics like precision, recall, F1-score, sensitivity, specificity, and accuracy.**

***Index Terms*— Chronic Kidney Disease(CKD), University of California Irvine(UCI), Random Forest(RF), Logistic Regression(LOG), Support Vector Machine(SVM), K-Nearest Neighbor(KNN), Naïve Bayes(NB)**

## I. INTRODUCTION

Due to high mortality rate, Chronic Kidney Disease (CKD) has received much attention. According to a survey made by World Health Organization (WHO), Chronic diseases have become a concern threatening developing countries[3].

*Grenze ID: 01.GIJET.10.1.217*

Another major concern regarding CKD is it does not show obvious symptoms in its early stages. Therefore, the disease may not be detected until the kidney loses about 25% of its function.[5] CKD is a kidney disorder, which can be treatable in its early stages, but it causes kidney failure in its late stages.
Early diagnosis and treatment of CKD will prevent its progression to kidney failure. The best way to treat chronic kidney disease is to diagnose it in the early stages, but discovering it in its late stages will lead to kidney failure, which requires continuous dialysis or kidney transplantation to maintain a normal life [6].
In the medical diagnosis of CKD, two medical tests are used to detect CKD. One by a blood test to check the glomerular filtrate [1] and another by a urine test to check albumin[2]. Due to the increasing number of CKD patients, the scarcity of specialist physicians, and the high costs of diagnosis and treatment, especially in developing countries, there is a need for computer-assisted diagnostics to help physicians and radiologists in supporting their diagnostic decisions. Machine Learning (ML) techniques can be used to facilitate medical diagnosis of CKD, indeed several studies have already used these techniques to improve clinical prediction in kidney-related disease and have reported an improved classification accuracy [7]–[10]. However, none of these studies identified the most important predictive attributes needed to improve diagnosis. If identified in CKD patients, such attributes could be used for computer-aided CKD screening and diagnostic tests.
The contributions of the proposed work are as follows.

1) Firstly KNN imputation is introduced to fill in the missing values in the data set. KNN is used to the data set, in with the diagnostic categories are unknown.
2) Secondly, Machine learning methods such as Logistic Regression (LOG), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NBs) classifier and Feed Forward Neural Network (FFNN) were used to establish CKD diagnostic models on the complete CKD datasets. The models with better performance were extracted for misjudgment analysis.

## II. Literature Review

Subasi et al [7] predicted that, how CKD can be diagnosed by using Machine Learning (ML) techniques. ML algorithms have showed to be an useful method in detection of abnormalities in different physiological data, and are, with a great success, employed in different classification tasks.
Polat et al [8] proposed Support Vector Machine (SVM) classification algorithm which was used to diagnose CKD. To diagnose the CKD, two essential types of feature selection methods namely, wrapper and filter approaches were chosen which reduce the dimension of CKD dataset.
Wibawa et al [9] developed machine learning method using ensemble learning and feature selection to improve the quality of CKD diagnosis. The CKD dataset was taken from UCI machine learning repository. UCI dataset contains 400 instances with 24 attributes including signs, symptoms and risk factors that might appear due to CKD.
Tazin et al [10] proposed the capability of the classification of SVM, DT, NBs and KNN algorithm, in analyzing the CKD dataset collected from UCI repository was investigated to predict the presence of kidney disease.
Gudeti et al [11] aimed is to differentiate the performance of various machine learning algorithms that are primarily based on its accuracy.
Qin et al [12] proposed a machine learning methodology for diagnosing CKD. The CKD data set was obtained from the University of California Irvine (UCI) machine learning repository, which has a large number of missing values.

## III. Problem Domain

Firstly K Nearest Neighbor (KNN) imputation was used to fill in the missing values, which selects several complete samples with the most similar measurements to process the missing data for each incomplete sample. Then the selected Samples were fed into classifiers such as Logistic Regression (LOG), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes (NBs) classifier and Feed Forward Neural Network (FFNN) for disease diagnosis. The CKD dataset was obtained from the University of California Irvine (UCI) machine learning repository, which has a large number of missing values. All classifiers showed promising results for diagnosing a dataset into CKD or a normal kidney(NOT CKD). Figure 1 shows the general structure of CKD diagnosis in this work.

### *A. Problem Definition*

### *B. Preprocessing*

The UCI dataset contained outliers and noise, so it must be cleaned up in a preprocessing stage. The preprocessing stage consist of

i) estimating missing values

ii) eliminating noise, such as outliers, normalization,

iii) checking of unbalanced data

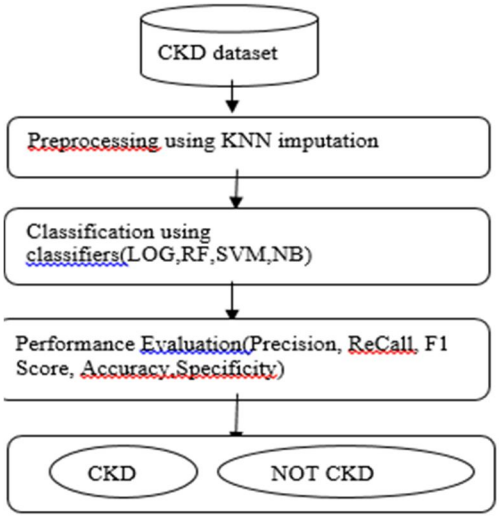


Fig. 1 Proposed Methodology

When patients are undergoing tests, some of the measurements may be missed and this causes missing values. The simplest method to handle missing values is to ignore the record, but it is illtrated for small dataset. KNN imputation is used in this study to compute missing values instead of removing records. The dataset used in this study contains numerical,categorical values. Each categorical (nominal) variable was coded to facilitate the processsing in a computer. For the values of rbc and pc, normal and abnormal were coded as 1 and 0, respectively. The missing values in the initial CKD data set were processed and filled initially after the categorical variables were encoded. This study uses KNN imputation, which determines for each sample with missing values with K complete samples have the lowest Euclidean distance. The category variable with the highest frequency in the corresponding variable in K complete samples is used to fill in the missing values, and for the numerical variables, the median of the corresponding variable in K full samples is used. People with similar physical conditions should have similar physiological data, which is why the KNN-based technique is used to fill in the missing values in physiological measurements. For instance, in healthy individuals, the physiological measurements should be steady within a specific range. When comparing the physiological data of individuals with similar degrees of a certain disease, identical results should be obtained. The values of K in this work were chosen as 3, 5, 7, 9 and 11. We thus possess comprehensive CKD data sets.

*C. Dataset*

The CKD dataset was collected from 400 patients from the University of California, Irvine Machine Learning Repository [13]. The dataset comprises 24 features divided into 11 numeric features and 13 categorical features, in addition to the class features, such as “ckd” and “notckd” for classification. Features include age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, and anemia. The diagnostic class contains two values: ckd and notckd. All features contained missing values except for the diagnostic feature. The dataset is unbalanced because it contains 250 cases of “ckd” class by 62.5% and 150 cases of “notckd” by 37.5%.It is worth mentioning that there is alarge number of missing values in the data

*D. Classification*

The machine learning classifiers used for this study are

1) Regression-based model: LOG

2) Tree-based model: RF
3) Decision plane-based model: SVM
4) Distance-based model: KNN
5) Probability-based model: NB
6) Neural network: FFNN
LOG uses linear regression as its foundation to determine a bias and the weight of each predictor.
By selecting training samples and predictors at random, RF produces a huge number of decision trees. A boundary that maximizes the difference between ckd and notckd is the goal of every decision tree's training process.
By creating a decision surface in a multidimensional space that includes the sample predictors, SVM separates several types of samples.
By computing the distances between the test sample and the training samples, KNN chooses the closest training samples. It then uses voting to establish the diagnostic category.
The number of ckd and notckd samples in each distinct measurement interval is used by the Naive Bayes classifier to determine the conditional probabilities of the sample under the interval.
Because of its intricate structure, FFNN can analyze non-linear relationships in data sets. The output layer and hidden layer both used the sigmoid activation function.
In the event that the sample distribution from the original data was retained, the entire data set was evenly split into four subgroups in order to assess model performance thoroughly.
Every subset was tested once for each of the aforementioned models, while other subsets were used for training. The final performance was determined by taking the average of all the subsets.

## IV. RESULTS

CKD was set to be positive and NotCKD to be negative in this investigation. The performance of the machine learning models was assessed and the specific results were displayed using the confusion matrix. The template of the confusion matrix is shown in Table 1.

TABLE 1. THE TEMPLATE OF CONFUSION MATRIX IN THIS STUDY

| Models at different values of K | Actual | Prediction | |
|---|---|---|---|
| Model | Class | ckd | notckd |
| | ckd | TP | FN |
| | notckd | FP | TN |

True Positive (TP) indicates the ckd samples were correctly diagnosed. False Negative (FN) indicates the ckd samples were incorrectly diagnosed. False Positive (FP) indicates the notckd samples were incorrectly diagnosed. True Negative(TN) indicates the notckd samples were correctly diagnosed. The model's performance was assessed using sensitivity, specificity, precision, recall, F1-score and accuracy.

$$\text{Recall(R)/ Sensitivity(Sen)} = \frac{\text{TP}}{\text{TP+FN}} \quad \text{....(1)}$$

$$\text{Specificity(Spe)} = \frac{\text{TN}}{\text{TN+FP}} \quad \text{....(2)}$$

$$\text{Precision(P)} = \frac{TP}{TP+FP} \quad \text{....(3)}$$

$$\text{F1} - \text{Score} = 2.\frac{\text{Precision}*\text{Recall}}{\text{Precision+Recall}} \quad \text{....(4)}$$

$$\text{Accuracy(Acc)} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad \text{....(5)}$$

These metrics results are evaluated using various classifiers under different K value as 11. Usually K defines the Eucledian mean distance. K can be taken as 3,5,7,9,11.. In our work the highest k value is used for the comparison of the performances.
The RandomForest Classifier shows a higher Sensitivity, Specificity,Precision,Recall, F1-Score and Accuracy, when compared to NB,KNN,FFNN,SVM,LOG. The following graph shows the comparison results.

TABLE II. COMPARISON OF THE MODELS' PERFORMANCES

| Classifiers | Evaluation Metrics (%) | | | | | |
|---|---|---|---|---|---|---|
| | **Sensitivity** | **Specificity** | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
| **NBs** | 88.134 | 89.871 | 88.471 | 88.134 | 88.302 | 88.585 |
| **KNN** | 88.879 | 91.541 | 89.714 | 88.879 | 89.296 | 91.436 |
| **FFNN** | 91.251 | 92.414 | 91.471 | 91.251 | 91.361 | 92.362 |
| **SVM** | 92.581 | 92.712 | 92.435 | 92.581 | 92.508 | 92.914 |
| **LOG** | 93.214 | 93.172 | 93.417 | 93.214 | 93.315 | 93.926 |
| **RF** | 94.514 | 94.421 | 94.832 | 94.514 | 94.673 | 94.932 |

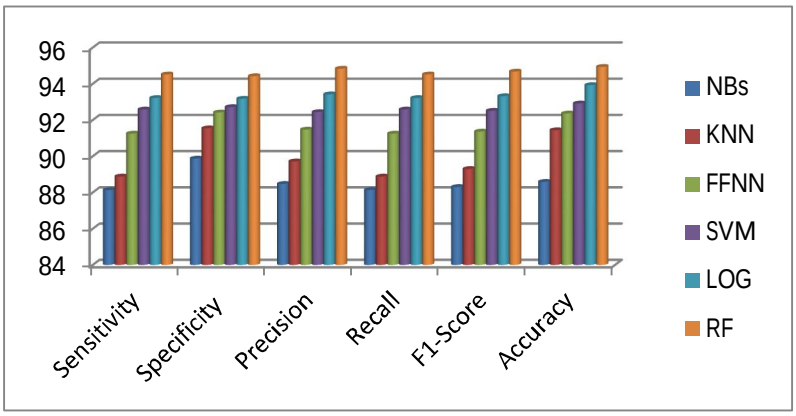


Fig 2. Comparison Results of Metrics of the Classifiers

## V. CONCLUSION

By imputing UCI dataset using KNN imputation and then feeding the output as input to the different classifiers the performance is evaluated. From the above table we can clearly state that Random Forest(RF) classifier showed highest performance when compared to the other classifiers. This shows that RF can be used to predict the CKD diseases in early stage which can save the life of huge number of people. In future ensemble/integrated model can be used to predict the disease in a better way.

## REFERENCES


[1] Arlet J. B., J. A. Ribeil, G. Chatellier, D. Eladari, S. de Seigneux, J. C. Souberbielle, G. Friedlander, M. de Montalembert, J. Pouchot, D. Prie, M. Courbebaisse, “Determination of the best method to estimate glomerular filtration rate from serum creatinine in adult patients with sickle cell disease: a prospective observational cohort study”, BMC Nephrology, vol. 13, no. 1, 2012.

[2] Sumida K., M. Z. Molnar, P. K. Potukuchi, F. Thomas, J. L. Lu, J. Jing, V. A. Ravel, M. Soohoo, C. M. Rhee, E. Streja, K. KalantarZadeh, C. P. Kovesdy, “Association of Slopes of Estimated Glomerular Filtration Rate With Post-End-Stage Renal Disease Mortality in Patients With Advanced Chronic Kidney Disease Transitioning to Dialysis”, Mayo ClinProc, vol. 91, no. 2, pp. 196-207, 2016.

[3] Zhang L., ``Prevalence of chronic kidney disease in China: A crosssectional survey," Lancet, vol. 379, pp. 815-822, Mar. 2012.

[4] Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, ``Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," J. Biomed. Informat., vol. 53, pp. 220-228, Feb. 2015.

[5] Cueto-Manzano M., L. Cortés-Sanabria, H. R. Martínez-Ramírez, E. Rojas-Campos, B. Gómez-Navarro, and M. Castillero-Manzano, ``Prevalence of chronic kidney disease in an adult population," Arch. Med. Res., vol. 45, no. 6, pp. 507-513, Aug. 2014.
[6] Hill, N.R., Fatoba, S.T., Oke, J.L., Hirst, J.A., O'Callaghan, C.A., Lasserson, D.S. and Hobbs, F.R., 2016. Global prevalence of chronic kidney disease–a systematic review and meta-analysis. PloS one, 11(7), pp.1-18.
[7] Subasi, E. Alickovic, and J. Kevric, ''Diagnosis of chronic kidney disease by using random forest,'' in Proc. CMBEBIH. Singapore: Springer, 2017, pp. 589–594.
[8] Polat H., H. D. Mehr, and A. Cetin, ''Diagnosis of chronic kidney disease based on support vector machine by feature selection methods,'' J. Med. Syst., vol. 41, no. 4, p. 55, Apr. 2017.
[9] Wibawa M. S., I. M. D. Maysanjaya, and I. M. A. W. Putra, "Boosted classifier and features selection for enhancing chronic kidney disease diagnose," in Proceedings of the 2017 5th international conference on cyber and IT service management (CITSM), pp. 1–6, IEEE, Denpasar, Indonesia, August 2017.
[10] Tazin, N., Sabab, S.A. and Chowdhury, M.T., 2016, Diagnosis of Chronic Kidney Disease using effective classification and feature selection technique. In 2016 International Conference on Medical Engineering, Health Informatics and Technology (MediTec) (pp. 1-6).
[11] Gudeti, B., Mishra, S., Malik, S., Fernandez, T.F., Tyagi, A.K. and Kumari, S., 2020, A Novel Approach to Predict Chronic Kidney Disease using Machine Learning Algorithms. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) ,pp. 1630-1635.
[12] Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C. and Chen, B., 2019. A machine learning methodology for diagnosing chronic kidney disease. IEEE Access, 8, pp.20991-21002.
[13] Dua D. and C. Graff, ``UCI machine learning repository," Ph.D. dissertation, School Inf. Comput. Sci., Univ. California, Oakland, CA, USA, 2017. [Online]. Available: http://archive.ics.uci.edu/ml.
[14] Hossain M. E., A. Khan, M. A. Moni, and S. Uddin, ''Use of electronic health data for disease prediction: A comprehensive literature review,'' IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 18, no. 2, pp. 745–758, Mar. 2021.
[15] Almansour, N.A., Syed, H.F., Khayat, N.R., Altheeb, R.K., Juri, R.E., Alhiyafi, J., Alrashed, S. and Olatunji, S.O., 2019. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in biology and medicine, 109, pp.101-111.
[16] Rubini, L.J. and Eswaran, P., , ''UCI chronic kidney disease,'' School Inf. Comput. Sci., Univ. California, Irvine, Irvine, CA, USA, 2015.