FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 843-847 | e-ISSN: 2347-470X

An Improved Deep Learning Approach for Prediction of The Chronic Kidney Disease

Akanksha[1] and Dr. Suganeshwari G[2]
[1]*Student, Department of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India, akanksha.2021@vitstudent.ac.in*
[2]*Assistant Professor, Department of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India, suganeshwari.g@vit.ac.in*

***Correspondence:** Akanksha; Email: akanksha.2021@vitstudent.ac.in

ABSTRACT- Kidney function is harmed by chronic kidney disease, leading to renal failure. Machine learning and data mining come in handy to detect kidney disease. Machine learning employs a variety of algorithms to make predictions and classify data. CT scans have been used to detect chronic renal disease. When CT scans are used to diagnose disease in the kidney, cross-infection occurs, and the results are delayed. The authors of the prior study developed a model for categorizing chronic renal illness utilizing multiple classification methods. A unique deep learning model is presented in this study for the early identification and prognosis of Chronic Kidney Disease (CKD). This study aims to build a neural network and evaluate its performance compared to other cutting-edge machine learning methods. Compared to the four different classifiers (K-Nearest Neighbor (KNN), Random Forest, Naive Bayes classifier, and probabilistic neural network), the suggested Deep neural model fared better by reaching higher accuracy. Nephrologists may find the proposed method helpful in the early detection of CKD.

General Terms: Data mining, Deep learning.
Keywords: Chronic kidney disease, Probabilistic neural network, K nearest neighbor, Support vector machine, logistic regression, decision tree algorithms, convolutional neural network.

ARTICLE INFORMATION
Author(s): Akanksha and Dr. Suganeshwari G.
Received: 12/06/2022; **Accepted**: 29/09/2022; **Published**: 18/10/2022;
e-ISSN: 2347-470X;
Paper Id: IJEER-RDEC6771;
Citation: 10.37391/IJEER.100414
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-10/ijeer-100414.html

Publisher's Note: FOREX Publication stays neutral with regard to Jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

We are aware of instances in which people worldwide are affected by various diseases due to environmental causes or lifestyle choices. These disorders must be foreseen or researched in advance, and the prognosis of the disease is critical in this instance. Machine learning and data mining algorithms can be used to provide predictions for the study of diseases.

Data mining aids in extracting and testing hidden data or data obtained as a dataset from huge medical field data sets of patients. In many circumstances, an early diagnostic methodology algorithm is required to establish kidney functionality, which is critical. A variety of classifiers were used to classify a CKD dataset in this study. The authors of [1-8] tried to perform a comparative analysis of various algorithms. They applied the algorithms like the Naïve Bayes Classifier and K-nearest neighbor, Decision tree, ANN, and PNN. K nearest neighbor had a 76.96 % accuracy rating, Decision tree had a 57.41 % rating, ANN had a 71.55 % accuracy rating, PNN had a 98.00 % rating, and Naive Bayes classifier had a 99.36 % rating. So, according to the conclusion section of the first research, the probabilistic neural network is the most efficient and effective means of forecasting chronic kidney diseases based on historical events and datasets.

In the proposed work, the structured and the unstructured data are integrated, and we propose a hybrid model. Compared to the other four classifiers (K-Nearest Neighbor (KNN), Random Forest, Naive Bayes classifier, and probabilistic neural network), the suggested Deep neural model fared better by reaching higher accuracy. Nephrologists may find the proposed method helpful in the early detection of CKD.

2. RELATED WORKS

Some widely implemented machine learning algorithms to diagnose kidney diseases were Naive Bayes, decision table (DT) (K-NN), random forest tree, K-nearest neighbor, and probabilistic neural networks. Only around five relatively wealthy nations, accounting for about 12% of the world's population, receive and endure care. As a result, detecting the early stages of CKD becomes necessary to improve the patient's speedy recovery and avoid fatal consequences [3].

Only approximately 100 developing countries treat nearly 20% of the world's population, accounting for almost half the worldwide population. Due to the high expense of dialysis or kidney transplantation, more than one million individuals die each year in 112 low-income countries from untreated renal insufficiency. The author of [1] discusses the four different algorithms employed. The implemented methods were the Radial basis function algorithm, probabilistic neural network, and multilayer perceptron support vector machine. Among the four algorithms, the probabilistic neural network is the most

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 843-847 | e-ISSN: 2347-470X

efficient and effective means of forecasting chronic kidney diseases based on historical events and datasets obtained.

The authors of [2] applied artificial neural networks for various algorithms. The KNN algorithm (K Nearest Neighbor) is also being employed, and the RFT algorithm is implemented. In this second study, the PNN (probabilistic neural network) method is the most effective and efficient technique for determining the algorithm that delivers the most effective relevant predictions. The authors in [3] studied the Bayes classifier decision tree and a random forest tree are three separate methods. The random forest tree algorithm is the most effective, efficient, and forward-looking among the three algorithms [3]. The decision tree obtained the highest result. The researchers in [4] applied algorithms like Logistic regression, decision trees, random forest trees, and convolutional neural networks to predict CKD. The convolutional neural network was the most effective approach employed among all the algorithms [4]. In contrast, support vector machines excelled over other algorithms with an accuracy rate of 95 % [5].

In the proposed work, we employ a decision tree, Random Forest, naive Bayes, and KNN to do an empirical study. Later we propose a deep learning method that integrates both structured and unstructured data to improve CKD's accuracy and early detection.

3. MATERIALS AND METHODS

3.1 Dataset

The chronic kidney dataset is available on the KAGGLE site, which most researchers use to search for studies in the area and try to discover the best strategy for predicting chronic kidney disease. The prediction is created based on several distinct parameters.

3.2 Methods

The following are the various algorithms used in the project:

3.2.1 Support Vector Machine

It is one of the most effective classification approaches for machine learning algorithms. In traditional learning approaches, SVM is based on structural risk minimization, which determines the hypothesis with the lowest possibility of errors. It is based on decreasing empirical risk and improving the performance of the learning set. As a result, quadratic optimization problems occur. It requires a wider range of patterns and a greater scale to handle complex classification problems. SVM is unaffected by the feature space's dimensionality.

The parameters being optimized are:
Kernel - values used included poly, rbf, linear and sigmoid.
C - values used were 0.1, 1, 10, 100, and 1000.
Gamma - values used were scale and auto.
Shrinking - values used were true and false.

3.2.2 Naïve Bayes

The Naive Bayes method is a simple and effective categorization system. It's usually used at the document level to categorize documents. It calculates the likelihood of various symptoms and categories in a given dataset. It primarily relies on feature-based processes, demands rapid and precise classification, and does not rely on large datasets.

$$P(A|B) = \frac{P(A,B))}{P(B)} \quad (1)$$

3.2.3 k-nearest neighbour

It detects the label category relevant to the training document in a similar test document. This method is used in KNN [2] to classify things into object-based classes. Only the function is estimated locally, and all computations differ until classification. It is used to calculate the Euclidean and Manhattan distances, which are used to predict ailments such as chronic renal disease.

Defining the knn parameters for grid search

knn_parameters_grid = {'n_neighbors': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], 'weights': ['uniform', 'distance'], 'algorithm': ['auto', 'ball_tree','kd_tree','brute'], 'n_jobs': [1, -1]}

The parameters being optimized are:
n_neighbours - values used are in the range of 1 to 10.
weights - values used were uniform, and distance
algorithm - values used were auto, ball_tree, kd_tree, and brute.
n_jobs - values used were 1 and -1.

3.2.4 Decision Trees

It detects the label category relevant to the training document in a similar test document. This method is used in KNN [2] to classify things into object-based classes. Only the function is estimated locally, and all computations differ until classification. It is used to calculate the Euclidean and Manhattan distances, which are used to predict ailments such as chronic renal disease.

$$E(S) = \sum_{i=1}^{C} -p_i log_2 p_i \quad (2)$$

Information gain in the decision tree is calculated as:

$$IG(D_P f) = I(D_P) - \frac{N_{left}}{N} I(D_{left}) - \frac{N_{right}}{N} I(D_{right}) \quad (3)$$

The parameters being optimized are:

Criterion - values used included Gini and entropy.
Splitter - values used were best and random.
Min_samples_leaf - values used were in the range of 1 to 5.
Max_features - values used were auto, sqrt, and log2.

3.2.5 Random Forest Tree Classifier

Its popularity is due to its ease of use and adaptability since it can handle classification and regression problems. Overfitting is less likely with decision trees since they tend to tightly fit all the samples inside training data. The random forest classifier is especially useful for estimating missing values because it retains accuracy even when a portion of the data is missing, thanks to feature bagging.

The parameters being optimized are:
Criterion - values used included Gini and entropy.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 843-847 | e-ISSN: 2347-470X

N_estimators - values used were multiples of ten from 10 to 100.
Min_samples_split - values used were in the range of 1.0 to 5.
Max_features - values used were auto, sqrt, and log2.

Figure 1: Diseased Kidney

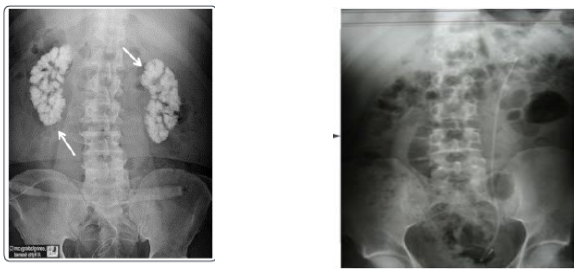

Figure 2: Non- diseased Kidney

4. PROPOSED SYSTEM

It consists of four stages which are explained in *figure 3*.

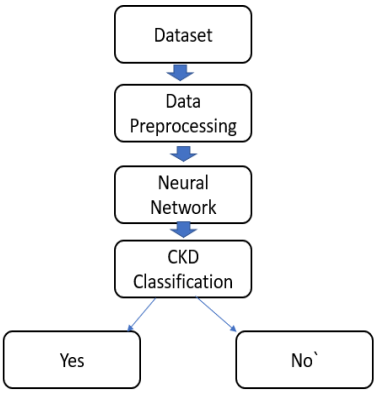

Figure 3: The architecture of the proposed method.

To develop a hybrid model, healthcare data can be in any format, such as a structured dataset or an image dataset, so in the proposed method, we employ an image dataset comprising diseased and non-diseased kidney x-ray pictures, as shown in *figures 1 and 2*.

4.1 Data Pre-processing

Data preprocessing is necessary before model development to remove a dataset's undesired noise and outliers. The dataset is then checked for null values. The dataset was divided into 70% for training and 30% for testing. KNN Imputation is performed to impute missing/NA values. Feature Scaling and Normalization are used to accelerate an algorithm's convergence and reduce training time.

4.2 For the Structured Dataset

STEPS

- Dataset collection
- Data preprocessing
- Data implementation of various algorithms
- Evaluation of the accuracy of the various algorithms
- Output of prediction

4.3 For Non-Structured Data

STEPS

- The annotated and unlabeled images are applied to the three separate convolutional layers.
- The labeled predicted pictures are obtained after the implementation and going through the convolutional layers.
- A probability map is created to determine if the patient has CKD.

Application of the CNN model on the image dataset and thus finding the accuracy of the model that how much the model can be efficient upon the application of the CNN model as shown in *figure 4*.

```
model = Sequential()
model.add(Conv2D(32,kernel_size=(3,3),activation='relu',input_shape=(224,224,3)))
model.add(Conv2D(64,(3,3),activation='relu'))
model.add(MaxPooling2D(pool_size=(2,2)))
model.add(Dropout(0.25))
```

```
model.add(Flatten())
model.add(Dense(64,activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(1,activation='sigmoid'))
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 222, 222, 32)      896

conv2d_1 (Conv2D)            (None, 220, 220, 64)      18496

max_pooling2d (MaxPooling2D  (None, 110, 110, 64)     0
)

dropout (Dropout)            (None, 110, 110, 64)      0

conv2d_2 (Conv2D)            (None, 108, 108, 64)      36928

max_pooling2d_1 (MaxPooling  (None, 54, 54, 64)       0
2D)

dropout_1 (Dropout)          (None, 54, 54, 64)        0

conv2d_3 (Conv2D)            (None, 52, 52, 128)       73856

max_pooling2d_2 (MaxPooling  (None, 26, 26, 128)      0
2D)

dropout_2 (Dropout)          (None, 26, 26, 128)       0

flatten (Flatten)            (None, 86528)             0
```

Figure 4: Model summary of the CNN model

FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 843-847 | e-ISSN: 2347-470X

5. RESULTS AND ANALYSIS

Upon implementing the various algorithms, the accuracy score encountered in the proposed model is that the values differ from the vary great range, thus making one of the algorithms the most efficient and, therefore, the most effective. *Table 1* shows the comparative study of various algorithms in terms of evaluation metrics.

Table 1: Comparative study of all algorithms

Classification Algorithm	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)
Support Vector Machine	95.8	94.7	96.4	95.6
K Nearest Neighbor	92.5	93.0	91.4	92.1
Decision Tree	93.3	89.5	96.2	92.7
Random forest	97.5	94.7	99.1	97.3

Figure 3 shows the plot of the various algorithms. *Figure 4* shows the true positive and true negatives of different machine learning algorithms.

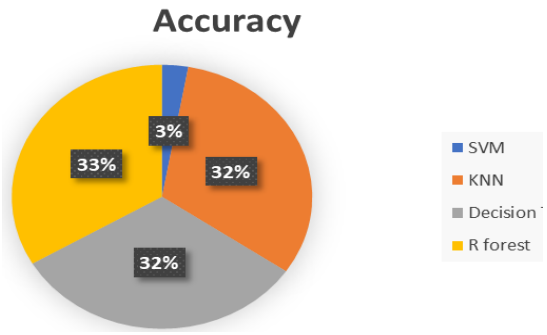

Figure 5: Accuracy of different algorithms

From *table 1* and *figures 5 and 6*, we infer that Random Forest has the highest F1-score (97.3%), accuracy (97.5%), recall (94.7%), and precision (100%) of all the other algorithms. To increase the fitting performance, certain modifications can be made to the metrics recorded, which are unreasonably very good. The provided dataset is quite small (only 500 records). Either a larger dataset must be used, or we can generate extra training examples using data augmentation (often used for deep learning).

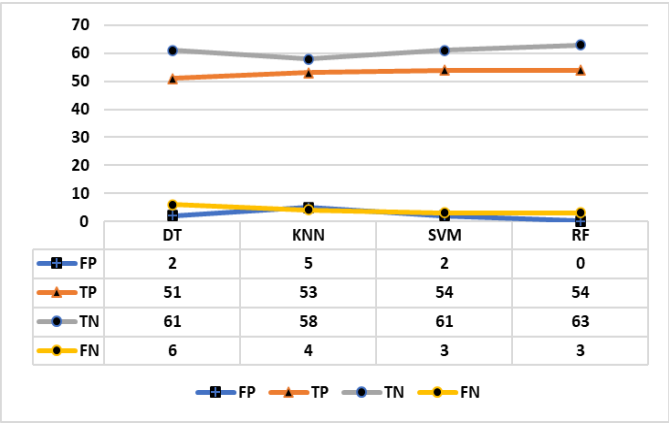

Figure 4: Comparison of True Positives and False positive values

- Improve data preprocessing by using min-max scaling, for example.
- Reduce the dimensions using PCA.

Although the data distribution has adequately covered the entire CKD domain, general characteristics like hunger, anemia, and pedal edema are biased in favor of CKD. Using this data set, it is simple to make an accurate prediction. However, generally speaking, it might result in false positives, as seen in *table 1* recall column. Because of the stage at which they manifest in the patient, some traits have a lower correlation than others when their medical value is taken into account. The accuracy of the models is greatly affected by the training process. Neural networks can be used to select the features, and then the model can be trained using tree structures to obtain higher accuracies.

6. CONCLUSION

Chronic kidney disease (CKD) affects over 14% of the world's population, and being able to predict it with 100% overall accuracy allows individuals to detect it early and receive treatment with the least amount of expense and risk. Effective feature engineering aids in reducing the number of features required for the prediction algorithm, in turn reducing the number of medical tests that must be performed. In summary, this work demonstrated the viability of using machine learning to assess the prognosis of CKD based on readily available data. In this investigation, random forest, naive Bayes, and logistic regression all showed equivalent predictability to the CKD dataset. Additionally, the sensitivity scores of these ML models were higher, which might be helpful for patient screens. Future research will involve integrating the output of the machine learning algorithms and neural network models to improve the accuracy of the prediction of Chronic Kidney Disease.

7. ACKNOWLEDGMENTS

We are also immensely grateful for the comments on an earlier version of the manuscript

REFERENCES

[1] Chittora, P., Chaurasia, S., Chakrabarti, P., Kumawat, G., Chakrabarti, T., Leonowicz, Z., & Bolshev, V. Prediction of chronic kidney disease-a machine learning perspective. IEEE Access, 2021, 9, 17312-17334.

[2] Ilyas, H., Ali, S., Ponum, M., Hasan, O., Mahmood, M. T., Iftikhar, M., & Malik, M. H. Chronic kidney disease diagnosis using decision tree algorithms. BMC nephrology, 2021, 22(1), 1-11.

[3] Krishnamurthy, S., Ks, K., Dovgan, E., Luštrek, M., Gradišek Piletič, B., Srinivasan, K., & Syed-Abdul, S. Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan. In Healthcare , 2021, 9, No. 5, p. 546

[4] Nishat, M. M., Faisal, F., Dip, R. R., Nasrullah, S. M., Ahsan, R., Shikder, F., & Hoque, M. A. A comprehensive analysis on detecting chronic kidney disease by employing machine learning algorithms. EAI Endorsed Transactions on Pervasive Health and Technology, 2021, 7(29)

[5] Kumar, A., & Pathak, M. A. A machine learning model for early prediction of multiple diseases to cure lives. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021, 12(6), 4013-4023.

[6] Almustafa, K. M. Prediction of chronic kidney disease using different classification algorithms. Informatics in Medicine Unlocked, 2021, 24, 100631.

FOREX Publication
Open Access | Rapid and quality publishing

International Journal of
Electrical and Electronics Research (IJEER)
Research Article | Volume 10, Issue 4 | Pages 843-847 | e-ISSN: 2347-470X

[7] Khan, B., Naseem, R., Muhammad, F., Abbas, G., & Kim, S. An empirical evaluation of machine learning techniques for chronic kidney disease prophecy. IEEE Access, 2020, 8, 55012-55022.

[8] Rady, E. H. A., & Anwar, A. S. Prediction of kidney disease stages using data mining algorithms. Informatics in Medicine Unlocked, 2019, 15.

[9] V Venkataiah, M Nagaratna and Ramakanta Mohanty (2022), Application of Chaotic Increasing Linear Inertia Weight and Diversity Improved Particle Swarm Optimization to Predict Accurate Software Cost Estimation. IJEER 10(2), 154-160. DOI: 10.37391/IJEER.100218.

[10] Dr. K. Sasikala, Dr. J. Jayakumar, Dr. A. Senthil Kumar, Dr. Shanty Chacko, Dr. Hephzibah Jose Queen (2022), Regression Based Predictive Machine Learning Model for Pervasive Data Analysis in Power Systems. IJEER 10(3), 550-556. DOI: 10.37391/IJEER.100324.