CHRONIC KIDNEY DISEASE PREDICTION

Jeevan Babu Maddala[1], Assistant Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
M. Vanaja[2], **P. Satya**[3], **N. Harika**[4], **N. Dinesh**[5]
[2,3,4,5] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
**jeevan@vvit.net[1], vanajamiddiboina.2002@gmail.com[2],
pulivarthisatya18@gmail.com[3], harikanarendra97@gmail.com[4],
neelamdinesh1912@gmail.com[5]**

DOI:10:48047/IJFANS/V11/I12/177

Abstract

Chronic kidney disease (CKD) is a global health issue with a high mortality rate and it is the root cause of many other diseases. Patients fail to recognise the disease as there aren't any obvious signs in the beginning. Early symptom identification is essential for providing effective treatment. To solve this problem, CKD was predicted using machine learning algorithms. In this study, CKD was predicted using convolutional neural networks (CNN) and long short-term memory (LSTM). We examine the accuracy, precision, F1-score, and recall of CNN and LSTM to see which performs better. The UCI Machine Learning repository is where the dataset was gathered.

Keywords: Chronic Kidney Disease, Classification, Convolutional Neural Network, Deep Learning, Health informatics, Long Short-Term Memory.

1. Introduction

Chronic kidney disease has one of the non-communicable illnesses with the quickest rate of growth (CKD). Every year, millions of individuals pass away; more than 10percent of the global population is impacted. Acute Renal Failure, Acute Nephritic Syndrome, and Chronic Kidney Disease are different types of the various kidney problems. Because kidney disease develops gradually over time, It is mentioned as a "chronic" disease. This illness increases the risk of cardiovascular disease, diabetes, and hypertension. So maintaining good kidney function is essential for overall health. The people who are aware of CKD at primary stage are only about 5%. This is why multiple CKD prediction models have been available. The popular machine learning methods are traditional methods like Logistic Regression (LR), Support Vector Machine(SVM) and K-Nearest Neighbours. Using machine learning methods, particularly deep learning, has gained popularity recently. Deep learning methods can identify certain patterns and extract relevant features from the dataset. These techniques can quickly and accurately identify CKD cases. In this context, Convolutional Neural Networks (CNNs) became popular model for classification.

In this study, we proposed CNN and LSTM are two deep learning models to predict CKD at an early stage. The models are composed of Using several nonlinear activation layers. These layers are trained to collaborate in a problem-solving manner. The main objective of this study is to identify patients who may be at risk of developing CKD.

2. Literature Survey

In this paper [1], they proposed three machine learning algorithms as SupportVector Machine, Logistic Regression and K-Nearest Neighbour. The CKD dataset from UCI with 25 attributes is used. The accuracy achieved was 99% for SVM, 77% for LR and 78% for KNN. They concluded that SVM predicts Chronic Kidney Disease better than LR and KNN.

In this paper [2], classification algorithms such as Naïve Bayes and Support Vector Machine are used to predict kidney diseases. A synthetic kidney function test (KFT) dataset was used. Four types of kidney disorders are classified. SVM has high accuracy of 76% when compared to Naïve Baye's whose accuracy is 70%.

In the paper [3], they focused on using CKD data from UCI with 12 best attributes. For this Ant Colony Optimization is used as feature selection method. The authors implemented Support Vector Machine (SVM) classifier and obtained an accuracy of about 96%.

In this work[4], they used three Machine Learning algorithms. The dataset used contains 400 records with 14 attributes. Logistic Regression obtained the highest accuracy of 97% while the accuracies are 71.25% for KNN and 96.25% for Decision Tree.

In the Paper [5], the importance of features is identified in the prediction of CKD. In this paper, a Deep Neural Network model was proposed with an accuracy of 98%. The dataset is collected from General Hospital in Gashua. The authors selected 10 attributes from the dataset.

In paper [6], they implemented an artificial neural network (ANN) which outperforms support vector machine (SVM). From UCI the dataset is collected. The accuracy is 99.75% for ANN and 97.75% for SVM.

In paper [7], they proposed the deep learning algorithms CNN, ANN[16], and LSTM as three optimised versions as well as traditional CNN[17], ANN, and LSTM models to predict CKD at the primary stage. They achieved accuracies of optimized CNN, ANN and LSTM are 98.75%, 96.25%, and 98.5%, respectively where as the achieved accuracies of CNN, ANN and LSTM are 92.71%, 90.43%, and 88.51% respectively.

In this work [8], Deep neural model was proposed by the authors which outperformed other five models by obtaining 100% accuracy. The accuracies of five models SVM, KNN, Random Forest, Decision Tree[13] and logistic regression are 92%, 92%, 97%, 95%, and 99% respectively.

In Research Paper [9], a multi-layer perceptron classifier is proposed to diagnose chronic kidney disease (CKD) using UCI dataset. They implemented a Deep Neural Network[14] model that achieves 100% accuracy. The accuracies of Random Forest[15], Decision trees, logistic regression and SVM are 96%, 96%, 92% and 81% respectively.

In this Paper [10], performance assessment of seven DL models including ANN, LSTM, Bi LSTM, GRU Bi GRU, simple RNN and Multi-Layer Perceptron is done to predict and diagnose CKD. The accuracy achieved is 99%, 85%, 88%, 85%, 89%, 96% and 96% respectively.

3. Problem Identification

Although there are various approaches for predicting CKD, most of the traditional methods take more time and may result in the omission of important features. It is crucial for us to predict CKD at early stage as the purpose of treatment is to prevent renal disease from progressing. To classify diseases, traditional machine learning approaches rely on expert-defined features. Deep Learning algorithms can learn relevant features from raw data automatically, making them more efficient and effective. This research paper aims to predict CKD using deep learning models. They are CNN and LSTM.

4. Proposed Methodology

The proposed system is developed with three phases which includes Data Collection, Model Building and Prediction. Datasets go through data preparation, this stage involves cleaning and changing the data into a format that can be easily utilised by the models. The data should be standardised, normalised, and handled correctly when there are missing values. After which the cleaned-up dataset is sent to training and testing. After training and testing, deep learning algorithms such as CNN and LSTM can be utilized for prediction. The proposed system procedure is shown in Figure 1.

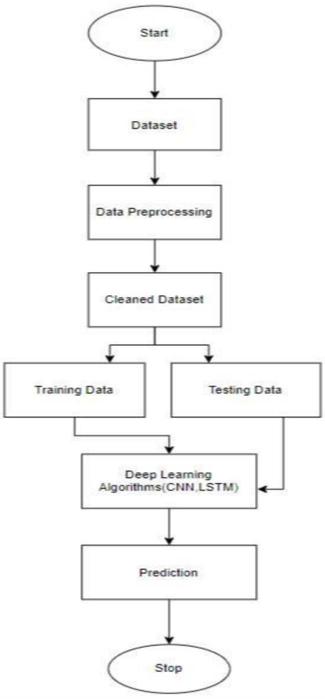
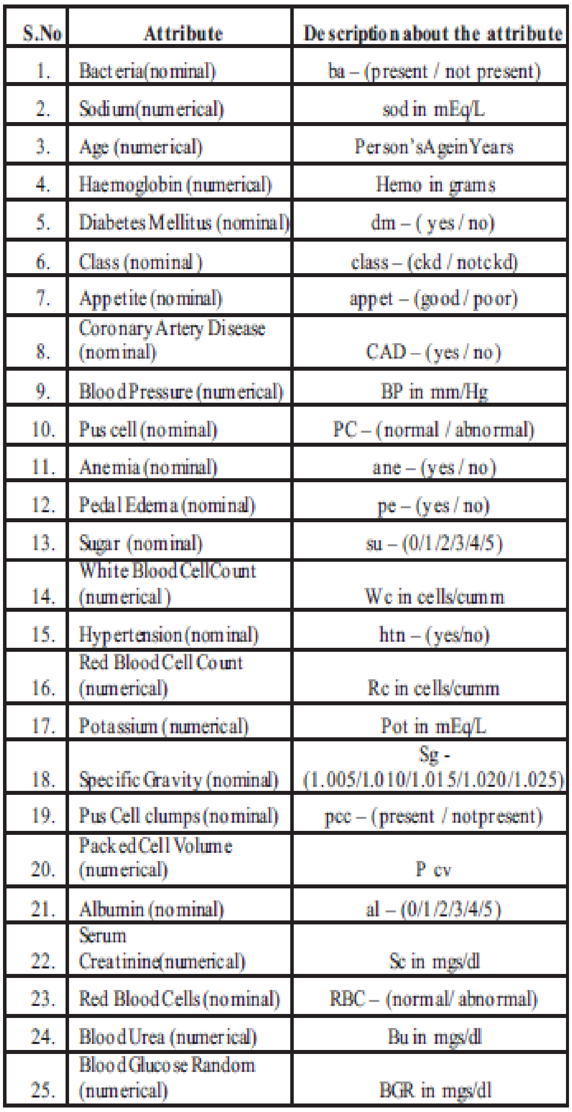


Figure 1: Process of Proposed System

4.1 Data Collection

From UCI Machine Learning Repository[18], Dataset of CKD was obtained.. There are 400 patients in this dataset. According to the examination of test results, 250 people have CKD and 150 do not. The dataset comprises twenty-five features, which are separated into fourteen numeric features and eleven categorical features, in addition to the category features of classification class, such as "ckd" and "notckd" for classification. The data features are depicted in Table 1.



S.No	Attribute	Description about the attribute
1.	Bacteria(nominal)	ba – (present / not present)
2.	Sodium(numerical)	sod in mEq/L
3.	Age (numerical)	Person'sAgeinYears
4.	Haemoglobin (numerical)	Hemo in grams
5.	Diabetes Mellitus (nominal)	dm – (yes / no)
6.	Class (nominal)	class – (ckd / notckd)
7.	Appetite (nominal)	appet – (good / poor)
8.	Coronary Artery Disease (nominal)	CAD – (yes / no)
9.	BloodPressure (numerical)	BP in mm/Hg
10.	Pus cell (nominal)	PC – (normal / abnormal)
11.	Anemia (nominal)	ane – (yes / no)
12.	Pedal Edema (nominal)	pe – (yes / no)
13.	Sugar (nominal)	su – (0/1/2/3/4/5)
14.	White BloodCellCount (numerical)	Wc in cells/cumm
15.	Hypertension (nominal)	htn – (yes/no)
16.	Red Blood Cell Count (numerical)	Rc in cells/cumm
17.	Potassium (numerical)	Pot in mEq/L
18.	Specific Gravity (nominal)	Sg - (1.005/1.010/1.015/1.020/1.025)
19.	Pus Cell clumps (nominal)	pcc – (present / notpresent)
20.	PackedCell Volume (numerical)	P cv
21.	Albumin (nominal)	al – (0/1/2/3/4/5)
22.	Serum Creatinine(numerical)	Sc in mgs/dl
23.	Red Blood Cells (nominal)	RBC – (normal/ abnormal)
24.	BloodUrea (numerical)	Bu in mgs/dl
25.	BloodGlucose Random (numerical)	BGR in mgs/dl

Table 1: shows the features of the UCI CKD data.

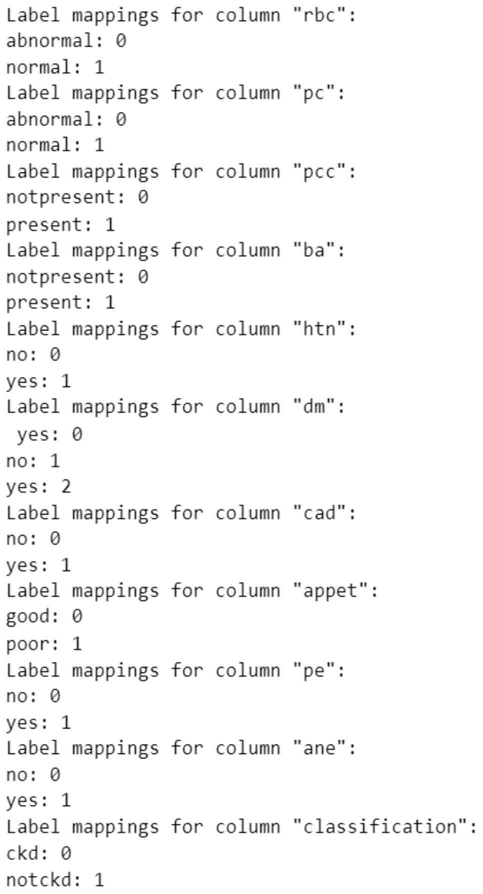
4.2 Data Preprocessing

4.2.1 Handling Missing Values

Missing data values are common in datasets. The prediction model's performance will suffer if the missing values are not handled correctly. Dropping missing data and filling missing values are two common approaches of handling missing values. Since there are both numeric and nominal data, the missing values are handled using mode imputation in this study.

4.2.2 Categorical Data Encoding

Label Encoder is used to normalise labels and convert non-numerical labels into numerical labels. Non numeric Label mappings are shown in Figure 2.



```
Label mappings for column "rbc":
abnormal: 0
normal: 1
Label mappings for column "pc":
abnormal: 0
normal: 1
Label mappings for column "pcc":
notpresent: 0
present: 1
Label mappings for column "ba":
notpresent: 0
present: 1
Label mappings for column "htn":
no: 0
yes: 1
Label mappings for column "dm":
 yes: 0
no: 1
yes: 2
Label mappings for column "cad":
no: 0
yes: 1
Label mappings for column "appet":
good: 0
poor: 1
Label mappings for column "pe":
no: 0
yes: 1
Label mappings for column "ane":
no: 0
yes: 1
Label mappings for column "classification":
ckd: 0
notckd: 1
```

Figure 2: Label Mappings for categorical data

4.2.3 Data Transformation

Data Scaling is done for numerical features before fitting the data to any model. There are various scaling methods, and Standard Scalar Normalization has been applied in this study. The data is scaled between −1 to +1. Values for a feature are normalised based on the mean and standard deviation. It is as follows:

$$Z = (a - \mu)/\sigma$$

where z is Z-score, a is feature value, μ is mean value and σ is standard deviation.

4.3 Classification

Training and testing datasets have now been created from the cleaned dataset. 20% of the cleaned dataset is used as the testing dataset to gather predictions, while 80% of the cleaned dataset is utilised as the training dataset to train the CNN and LSTM models. The

CNN model is used for classification of CKD because of better performance compared to LSTM.

5. Implementation

A system was developed using two models such as CNN and LSTM in python language.

5.1 Convolutional Neural Network (CNN)

For working with numerical data, the Convolutional Neural Network (CNN) architecture is the best approach. The key features are extracted by several filters when the numerical data is passed through convolutional layers. To extract features from input data, maxpooling and convolutional layers (Con2D) are utilised. The extracted features are sent to the fully connected layers to make a prediction using activation functions. Our architecture's major layers are the Covolution layer, the Pooling layer, and the Fully - connected layer. The below Figure 3 is the Working of CNN Architecture [11].

Figure 3: CNN Architecture to predict CKD

Convolution Layer:

In this study, we used 2D Convolutional layers, which can learn patterns in the data across both the feature and sample axes.

Max pooling:

Max pooling selects the maximum value within a certain spatial window that by down-sampling reduces the spatial dimensions of the feature maps.

Fully Connected Layer (Dense):

The pooling layers' flattened output is processed and transformed into a set of high-level features relevant to distinguish between ckd and not ckd. The dense layer updates the weights of input data to generate a prediction, the fully connected layer applies a series of biases and weights to the input data.

Output Layer:

The softmax function is employed as the activation function for the output layer. Softmax is an activation function that predicts a multinomial probability distribution. The final prediction is generated by the output layer.

5.2 Long Short-Term Memory (LSTM)

For numeric dataset, LSTM is used to model the temporal relationships between the various features. We designed an LSTM model that takes the pre-processed input data and produces an output that predicts whether or not a patient has CKD. The model includes one LSTM layer, followed by fully connected layers that output a binary classification result. The layers used in lstm for ckd prediction are shown in figure 4. In this model, the input is fed into an LSTM layer, which is a recurrent neural network type. Three gates and a memory cell constitute the LSTM's basic architecture.

Preprocessed Input
|
LSTM layer
|
Dropout layer
|
Dense Layer
|
Output

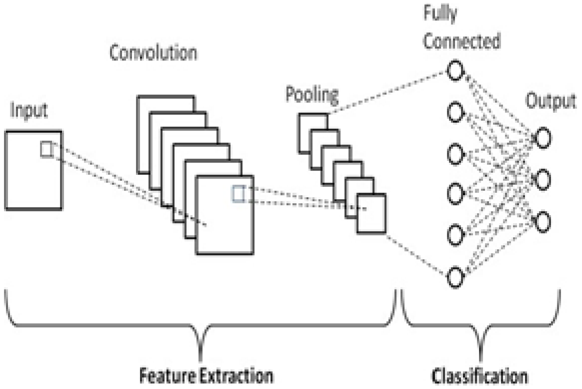


Figure 4: LSTM architecture

The memory cell consists of a cell state and hidden state memory to enable the network to learn long-term dependencies. Information is stored through time steps without substantial changes in the cell state (c), which serves as long-term memory. The hidden state (h) serves as a short-term and is frequently updated. At time-step t, c_t and represent the cell state and hidden state, respectively.. The three gates are defined below[12]:

$i_t = \sigma(W_i * X_t + U_i * h_{t-1} + b_i)$ (1)

$f_t = \sigma(W_f * X_t + U_f * h_{t-1} + b_f)$ (2)

$o_t = \sigma(W_o * X_o + U_o * h_{t-1} + b_o)$ (3)

where i_t is the input gate, f_t is the forget gate and o_t is the output gate. X_t is the cell input. σ is the sigmoid function. W, U, and b are the weight matrix, recurrent weight matrix, and bias the relevant gates' vector. The input gate controls the quantity of new data added to the cell state. The forget gate regulates how much data is taken out of the cell state. How much information from the cell state is sent to the next hidden state is controlled by the output gate. For each input, the cell state and hidden state are modified as

$g_t = \phi(W_g * X_t + U_g * h_{t-1} + b_g)$ (4)

$$c_t = f_t * c_{t-1} + i_t * g_t \qquad (5)$$

$$h_t = o_t * \phi(c_t) \qquad (6)$$

where * and + represent element-wise multiplication and addition. g_t is the candidate value that represents the new information that could be added to the cell state. and ϕ represents activation function. We are using ReLU in this model. The below Figure 5 represents LSTM cell architecture[12].

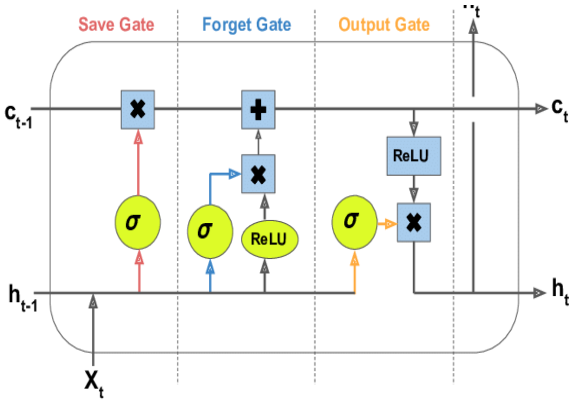


Figure 5: LSTM cell architecture

To avoid overfitting, a dropout layer with a dropout rate of 0.5 is placed to the LSTM layer. A single unit with softmax activation makes the final prediction in the output layer.

6. Results & Conclusion

The proposed model (CNN best of two) is able to classify CKD and not CKD with classification accuracy of 98.75% for 80%- 20% training-testing partition. After training of two models (CNN and LSTM), accuracy is compared. High accuracy was attained by CNN as shown in Table2. So, it can be concluded that CNN is used to predict Chronic Kidney Disease.

Model	CNN	LSTM
Accuracy	98.75%	96.25%
Precision	98.6%	96.0%
Recall	98.8%	96.6%
F1-score	98.7%	96.2%

Table 2: Comparison of Models

Figure 6 represents the Confusion Matrix obtained for CNN. Figure 7 shows Confusion Matrix of LSTM. From this confusion matrix the True Positive, True Negative, False Positive and False Negative values are predicted. Figure 8 shows the comparision graph of two algorithms. Figure9 shows the result obtained for the test data.

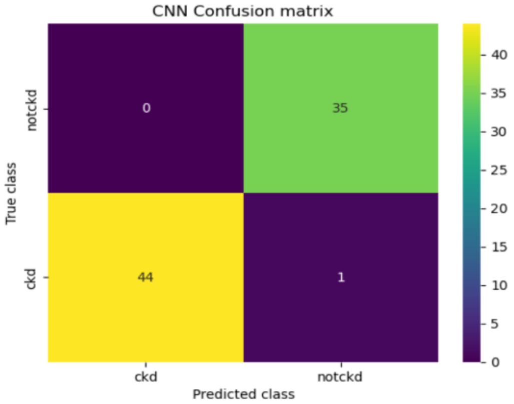


Figure 6: Confusion Matrix Of CNN

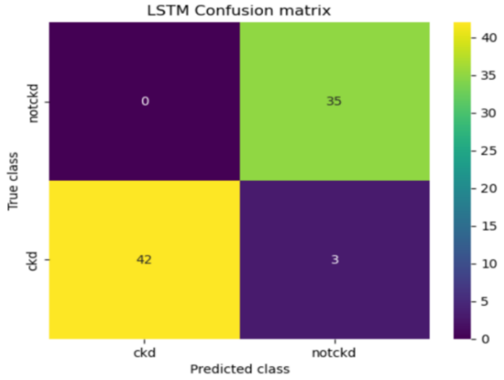


Figure 7: Confusion Matrix Of LSTM

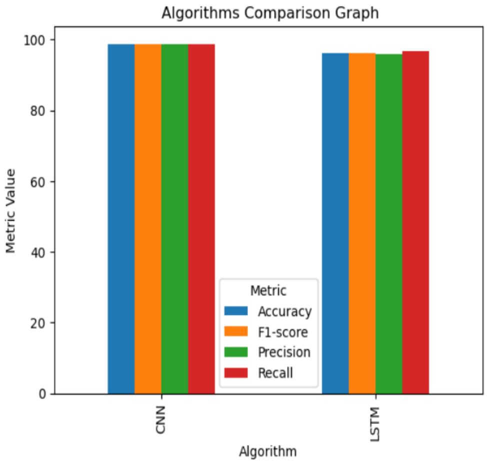
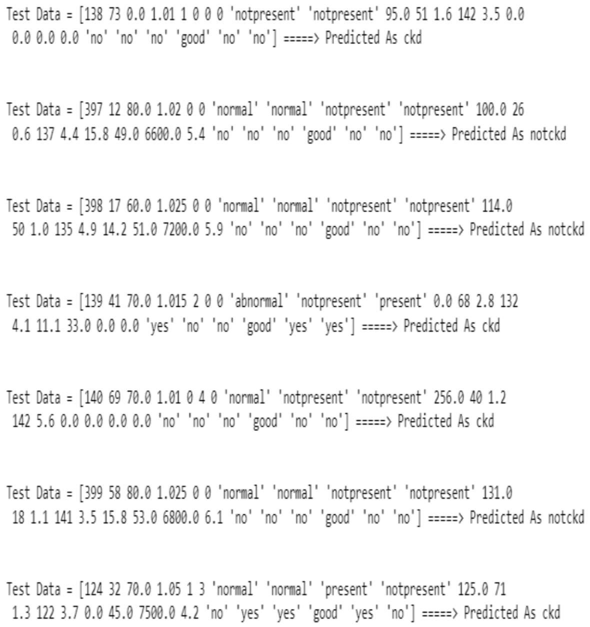


Figure 8: Comparison Graph



```
Test Data = [138 73 0.0 1.01 1 0 0 0 'notpresent' 'notpresent' 95.0 51 1.6 142 3.5 0.0
 0.0 0.0 0.0 'no' 'no' 'no' 'good' 'no' 'no'] =====> Predicted As ckd


Test Data = [397 12 80.0 1.02 0 0 'normal' 'normal' 'notpresent' 'notpresent' 100.0 26
 0.6 137 4.4 15.8 49.0 6600.0 5.4 'no' 'no' 'no' 'good' 'no' 'no'] =====> Predicted As notckd


Test Data = [398 17 60.0 1.025 0 0 'normal' 'normal' 'notpresent' 'notpresent' 114.0
 50 1.0 135 4.9 14.2 51.0 7200.0 5.9 'no' 'no' 'no' 'good' 'no' 'no'] =====> Predicted As notckd


Test Data = [139 41 70.0 1.015 2 0 0 'abnormal' 'notpresent' 'present' 0.0 68 2.8 132
 4.1 11.1 33.0 0.0 0.0 'yes' 'no' 'no' 'good' 'yes' 'yes'] =====> Predicted As ckd


Test Data = [140 69 70.0 1.01 0 4 0 'normal' 'notpresent' 'notpresent' 256.0 40 1.2
 142 5.6 0.0 0.0 0.0 0.0 'no' 'no' 'no' 'good' 'no' 'no'] =====> Predicted As ckd


Test Data = [399 58 80.0 1.025 0 0 'normal' 'normal' 'notpresent' 'notpresent' 131.0
 18 1.1 141 3.5 15.8 53.0 6800.0 6.1 'no' 'no' 'no' 'good' 'no' 'no'] =====> Predicted As notckd


Test Data = [124 32 70.0 1.05 1 3 'normal' 'normal' 'present' 'notpresent' 125.0 71
 1.3 122 3.7 0.0 45.0 7500.0 4.2 'no' 'yes' 'yes' 'good' 'yes' 'no'] =====> Predicted As ckd
```

Figure 9: Predicting CKD or not CKD for test data

7. Limitations & Future Scope

The limitation of proposed model is that it is trained with small datasets. In the future, large amounts of CKD data should be gathered in order to enhance the model performance. The future scope of this study is to implement deep learning models for predicting various stages of CKD. And Feature selection methods can be applied on dataset before training.

References

[1] Gudeti, B., Mishra, S., Malik, S., Fernandez, T. F., Tyagi, A. K., & Kumari, S. (2020, November). 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). IEEE.

[2] Vijayarani, S., & Dhayanand, S. (2015). Data mining classification algorithms for kidney disease prediction. *Int J Cybernetics Inform*, *4*(4), 13-25.

[3] Scholar, P. G. (2018). Chronic kidney disease prediction using machine learning. International Journal of Computer Science and Information Security (IJCSIS), 16(4).

[4] Gazi Mohammed Ifraz, Muhammad Hasnath Rashid, Tahia Tazin, Sami Bourouis, Mohammad Monirujjaman Khan, "Comparative Analysis for Prediction of Kidney Disease Using Intelligent Machine Learning Methods", *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 6141470, 10 pages, 2021. https://doi.org/10.1155/2021/6141470

[5] Iliyas, I. I., Saidu, I. R., Dauda, A. B., & Tasiu, S. (2020). Prediction of Chronic Kidney Disease Using Deep Neural Network. *arXiv preprint arXiv:2012.12089*.

[6] Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, & 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*, 101-111.

[7] Mondol, C., Shamrat, F. J. M., Hasan, M. R., Alam, S., Ghosh, P., Tasnim, Z., ... & Ibrahim, S. M. (2022). Early Prediction of Chronic Kidney Disease: A Comprehensive Performance Analysis of Deep Learning Models. Algorithms, 15(9), 308.

[8] Singh, V., Asari, V. K., & Rajasekaran, R. (2022). A deep neural network for early detection and prediction of chronic kidney disease. Diagnostics, 12(1), 116.

[9] Rahul Sawhney, Aabha Malik, Shilpi Sharma, Vipul Narayan. A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease. Decision Analytics Journal, Volume 6, 2023, 100169, ISSN 2772-6622 https://doi.org/10.1016/j.dajour.2023.100169.

[10] Akter, S., Habib, A., Islam, M. A., Hossen, M. S., Fahim, W. A., Sarkar, P. R., & Ahmed, M. (2021). Comprehensive performance assessment of deep learning models in early prediction and risk identification of chronic kidney disease. *IEEE Access*, *9*, 165184-165206.

[11] Phung, V. H., & Rhee, E. J. (2019). A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. Applied Sciences, 9(21), 4500.

[12] Grover, H., Alladi, T., Chamola, V., Singh, D., & Choo, K. K. R. (2021). Edge computing and deep learning enabled secure multitier network for internet of vehicles. IEEE Internet of Things Journal, 8(19), 14787-14796.

[13] Tekale, Siddheshwar, Pranjal Shingavi, Sukanya Wandhekar, and Ankit Chatorikar. "Prediction of chronic kidney disease using machine learning algorithm." *International Journal of Advanced Research in Computer and Communication Engineering* 7, no. 10 (2018): 92-96.

[14] Dutta, S., & Bandyopadhyay, S. K. (2020). Chronic kidney disease prediction using neural approach. *medRxiv*, 2020-06.

[15] Pasadana, I. A., Hartama, D., Zarlis, M., Sianipar, A. S., Munandar, A., Baeha, S., & Alam, A. R. M. (2019, August). Chronic kidney disease prediction by using different decision tree techniques. In *Journal of Physics: Conference Series* (Vol. 1255, No. 1, p. 012024). IOP Publishing.

[16] Dubey, G., Srivastava, Y., Verma, A., & Rai, S. (2021). Chronic Kidney Disease Prediction Using Artificial Neural Network. In *Proceedings of International Conference on Big*

Data, Machine Learning and their Applications: ICBMA 2019 (pp. 395-401). Springer Singapore.

[17] Navaneeth, B., & Suchetha, M. (2020). A dynamic pooling based convolutional neural network approach to detect chronic kidney disease. *Biomedical Signal Processing and Control*, *62*, 102068.

[18]http://www.ics.uci.edu/~mlearn/MLRepository.html. (2007).