International Journal of Modern Computation, Information, and Communication Technology
2022;5(2):52-62.
ISSN: 2581-5954

Machine Learning based Detection Technique to Predict the Survival of Patients with Chronic Kidney Diseases

Sofia Bobby J[1*], Annapoorani C L[1], Jerusha Renae A[1], Suruthi E T[1], Shineka K[1]

[1]*Department of Biomedical Engineering, Jerusalem College of Engineering, Chennai, Tamil Nadu, India.*

[*]*Corresponding author: sofiabobbyj@jerusalemengg.ac.in*

Abstract. There are currently many people all around the world who are suffering from chronic kidney infections. Today, everyone is attempting to be health-conscious, even though, owing to overwork and a hectic schedule, one only pays attention to one's health when symptoms appear. A few factors, for example, dietary habits, temperature, and expectations for daily luxuries, cause large numbers of people to be afflicted unexpectedly and without knowledge of their condition. Finding persistent kidney disease is often intrusive, costly, time-consuming, and dangerous. The main reasons why many people die without receiving care are especially in many developing countries since resources are few. As a result, early diagnosis and recognition of illness remains important, particularly in non-industrialized countries where illnesses are typically studied in late stages. However, if it does not show any symptoms at all, or if it does not show any disease-specific symptoms, it is very difficult to find &predict the disease type, detect and prevent such a disease, and this could lead to permanent health damage as well as the formation of new diseases, but machine learning can be a hope, as it is the best way for prediction and disease analysis. We will utilize data from CKD patients with 14 variables, as well as several machine learning approaches such as Decision Tree, SVM, and CNN model. To create an efficient machine learning model with the highest accuracy (by comparing several machine learning models) in predicting whether or not a person has CKD and, if so, how severe it is.

Keywords: Chronic Kidney Disease, machine learning, deep learning, sequential model, Neural network.

INTRODUCTION

Recent advancements in information technology during the previous ten years include cellular communication systems and machine learning, both of which are used in the field of healthcare. Various intelligent healthcare systems are being modeled with the use of big data and mobile computing devices to provide intellectual and expert services. Furthermore, the increase in medical data raises several challenges in terms of data handling, storage, and processing. Low-grade inflammation is increasingly recognized as a critical component of chronic kidney disease (CKD) [1]. There has been a challenging and significant advancement on the healthcare side, with CKD still being a crucial health problem that affects 10–15 percent of the population, and its pervasiveness is constantly expanding [2].

Because of its delicate nature, CKD is frequently misdiagnosed in its early stages. A person with CKD is more likely to develop heart disease. The early stages of CKD do not exhibit any significant symptoms, making it difficult to diagnose without the use of diagnostics like urine and blood tests [3]. When CKD is recognized in its early stages, preventative measures and improved treatment can be implemented to reduce the likelihood of dialysis or transplantation. According to one research, early diagnosis of CKD by nurses specializing in nephrology and primary care doctors can help slow the progression of the condition. Imaging methods are commonly used to detect the presence of CKD. However, due to the vast number of patients, it is impractical to examine everyone, and people who are at a higher risk of developing CKD will be advised to undertake comprehensive testing [4].

In existing work, they used the Convolutional Neural Networks (CNN) to detect & classify the CKD. But they selected the priority & nonpriority features based on heat map correlation using Seaborn library [5]. Correlation feature values may slightly confuse the users & researchers, because all values are in decimal with negative values

Received: 13.08.2022 Revised: 29.09.2022 Accepted: 12.10.2022
Licensed under a CC-BY 4.0 license | Copyright (c) by the authors

also that time, there is possibly to miss some priority features from dataset. They use the data split with 75% -25for training & testing [6].

In this study, several machine learning approaches were used to a dataset including information regarding patients' chronic kidney disease diagnoses. These approaches are Nave Bayes, Support Vector Machine, Decision Tree, Random Forest, Nearest Neighbor Algorithm, and CNN model. Predictions are made using twenty-five attributes from the data set. Some characteristics have been shown to be important in decision-making. The results obtained from the machine learning method performed comparatively better than previous techniques and may be useful for predicting CKD [7].

LITERATURE REVIEW

The many technologies of data mining (DM) models for forecasting cardiac disease are reviewed in this research. Data mining is useful in developing an intelligent model for medical systems to identify heart disease (HD) utilizing patient data sets that include risk factors linked with heart disease [8]. Medical professionals can assist patients by anticipating cardiac disease before it occurs. Using data mining technologies, enormous amounts of data from medical diagnoses are evaluated, and usable information known as knowledge is retrieved [9]. Mining is a way of studying enormous volumes of data to extract hidden and previously undiscovered patterns and correlations, as well as knowledge detection, to aid in better comprehension of medical data to avoid heart disease.

A unique ensemble learning strategy known as the "BBS method," which stands for Bagging, Boosting, and Stacking, was used to classify five UCI datasets from the field of bioinformatics. Experiments are carried out with Weka and Java Eclipse, and it has been empirically demonstrated that our approach provides superior accuracy with a lower root mean square error rate when employing the ensemble learning strategy. As a result, we conclude that our suggested ensemble learning technique is more suited for dealing with the classification problem in the bioinformatics sector. Such techniques can be employed effectively in relevant real-life categorization contexts [10].

The difficult tasks of picking important features from a vast amount of accessible information and detecting heart disease are carried out in this study. One of the most common pre-processing processes in classification issues is feature selection. A modified differential evolution (DE) method is employed to accomplish feature selection and optimization for cardiovascular disease [11].

Chronic renal disease is a deadly kidney disorder that can be avoided with early detection and sufficient safeguards, according to this paper [12]. Data mining of previously diagnosed patients' information ushered in a new era of medical progress. However, special procedures must be used to achieve a better result [13]. The competence of the classification algorithms Support Vector Machine, Decision Tree, Nave Bayes, and K-Nearest Neighbor in assessing the chronic kidney disease dataset gathered from the UCI repository was explored in this work to predict the existence of kidney disease [14]. The accuracy, Root Mean Squared Error, Mean Absolute Error, and Receiver Operating Characteristic curve of the data set were all calculated. Decision tree yields encouraging outcomes in this investigation [15].

PROPOSED METHODOLOGY

In general, dataset features are directly applied to the algorithms to get the performance of every classifier. In the existing system, they used the ANN model to predict the chronic kidney disease. But in our model, we are going to use feature Selection method (Random Forest). Initially we are going to call the machine learning models, based on the feature importance are calculated [16]. Those results are plotted like bar chart/graph with sorted like higher to lower. Based on results, we can decide which of the features are priority& nonpriority. In that lower-level value of feature vector are neglected from data frame. Once they are selected, apply these values to the classifier and their performance is noted. To get the good accuracy of classifier ML model. In other hand 3 methods of custom CNN sequential models was created as shown in Figure 2, and its results are compared with machine learning model results. Figure 1 shows the workflow diagram.

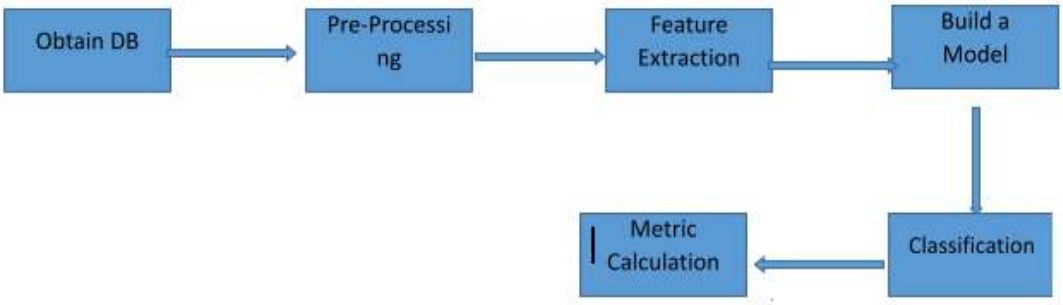

FIGURE 1. Workflow Diagram

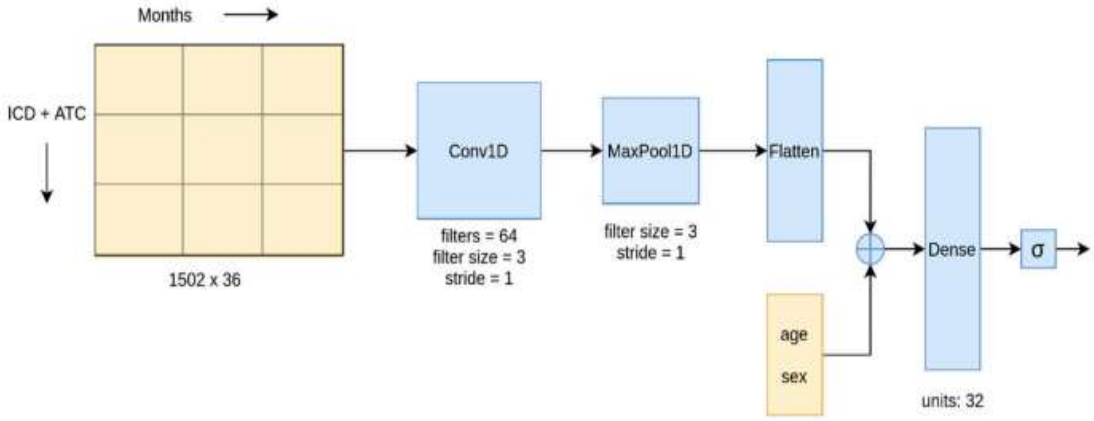

FIGURE 2. CNN Architecture

DEMERITS FOR EXISTING METHOD

- High complexity.
- Feature selection methods are not added
- Unable to control the learning process.

Initially we have 24 features are taken up for the machine learning classification. Before initiate the machine learning process, we are going to use visualize the data using plotly & Seaborn library. After that we used the classifiers like Support vector, K Nearest neighbour, decision tree classifier, Random Forest. Before passing the dataset to the classifier, entire data can be split into 2 parts as training and testing. 80% allocated for training & 20% allocated for testing. After the data split, training data is applied to the machine learning classifiers. Based on the training, a test data is taken up for validation or prediction i.e., to find the performance of the classifier. Figure 3 shows the general architecture.

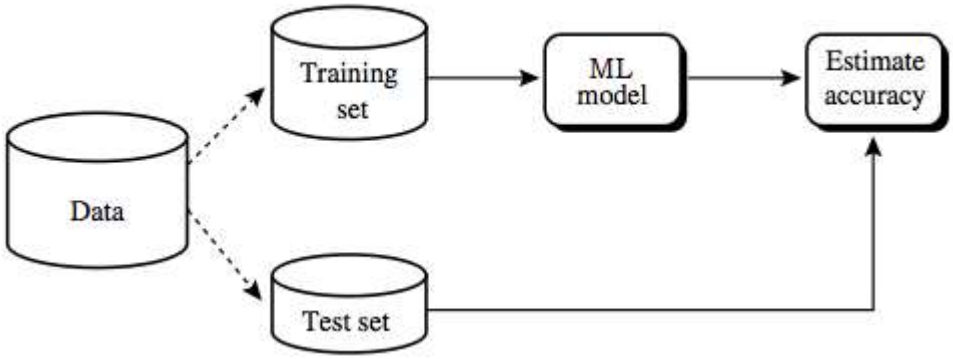

FIGURE 3. General Architecture

But in our proposed work, in feature selection, whose priority of their feature vector is gathered from RFE method, in that lower-level value of feature vector are neglected from data frame. Machine learning (ML) is an area of knowledge that has developed from Artificial Intelligence. Machine Learning can better be described by researcher Samuel in 1959 as a topic that allows the machine to learn while being specifically programmed. There are several kinds of machine learning classification methods. The main two of them are supervised and unsupervised instruction. Other than those mentioned above, is reinforcement learning, recommended systems, etc. The concepts of issues that can be discussed by machine learning methods are regression, classification and clustering.

The supervised machine learning job is undergoing two major phases, i.e. the training period and the testing period. During the training period, the graph is built and evaluated in the same way during testing stage. The possibility of using machine learning for decision-making in a variety of fields has a critical part to play in people's advancement. There are several various machine learning techniques, and the 3 main types under which they can be categorized are supervised, unsupervised, and reinforced.

Supervised learning is a concept, consisting of a goal to be anticipated among the given data set offered. Output data as well as the input set of variables, often referred to as training images, a mapping is made of which corresponds to the appropriate output. The training shall be accompanied by a series of sequential methods of implying and shall continue to obtain the necessary benchmark of precision. Several examples of supervised learning include Linear Regression, Logistic Regression, Decision Tree, and SVM [17]. Many disease detection systems are discussed for skin cancer, oral cancer, ECG signal classification and DNA classification in [18-21].

Unsupervised learning principles have no goal or outcome variable for making conclusions in trained data. Since the measurements provided to the learner are not numbered, there is no assessment of the quality of the production by the appropriate algorithm. It is widely used for clustering issues. Few of the examples that come under this category are Apriori Algorithm and K-Means. Reinforcing learning methods use previous activity to gain the best information and to turn it to correct business decisions. In this situation, the computer was subjected to a training environment, constantly implementing different techniques, to determine the behaviour would provide the maximum reward. Markov Decision Process is one such instance of improved learning.

RESULTS AND DISCUSSION

geting index value of important features

```
[ ]  Feature_selector_index = Feature_selector.get_support(indices=True)
     Feature_selector_index

     array([ 2,  4,  5,  6,  7,  8,  9, 14, 15, 16, 17, 18, 19, 20, 21])
```

getting seleceted column names

```
[ ]  Data_Features=X_feat_data_1.columns

     Feature_selector_col = Data_Features[Feature_selector_index]
     Feature_selector_col

     Index(['specific_gravity', 'sugar', 'red_blood_cells', 'pus_cell',
            'pus_cell_clumps', 'bacteria', 'blood_glucose_random', 'haemoglobin',
            'packed_cell_volume', 'white_blood_cell_count', 'red_blood_cell_count',
            'hypertension', 'diabetes_mellitus', 'coronary_artery_disease',
            'appetite'],
```

FIGURE 4. Feature Selection results from RFE (Recursive Feature Elimination)

After applying the Feature selection (RFE) as shown in Figure 4, less priority features are removed. Only 15 features (which are highly important). Then priority features are again applied to the machine learning classifiers to get the performance.

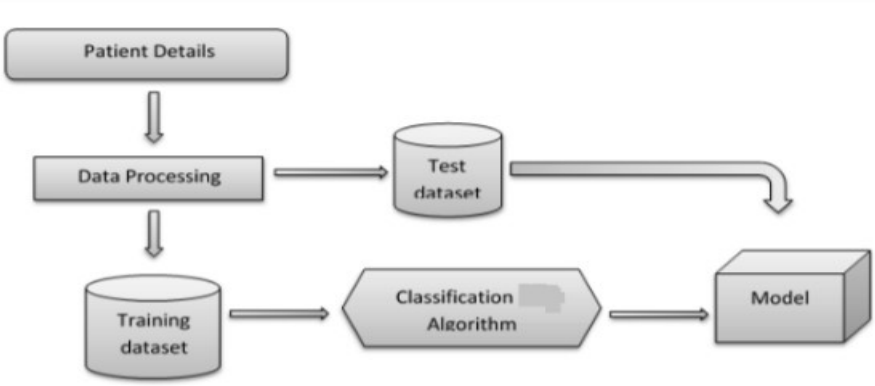

Figure 5. Flow of the Proposed System

Figure 5 shows the workflow of the system. After that, the collection of machine learning performance compared with 3different custom CNN models. This is identified the best classifier to detect & classify the phishing website. The complete implementation can be done through Google Colab (Python-Jupyter Notebook).

MERITS (PROPOSED)

- Ability to generate new features from limited series of features located in the training dataset.
- Easy to manage a large set of experiments
- Less computational time.

After removing the less priority features are removed from our existing dataset as shown in Figure 6.

geting index value of important features

```
[ ]  Feature_selector_index = Feature_selector.get_support(indices=True)
     Feature_selector_index
```

```
array([ 2,  4,  5,  6,  7,  8,  9, 14, 15, 16, 17, 18, 19, 20, 21])
```

getting seleceted column names

```
[ ]  Data_Features=X_feat_data_1.columns
```

```
Feature_selector_col = Data_Features[Feature_selector_index]
Feature_selector_col
```

```
Index(['specific_gravity', 'sugar', 'red_blood_cells', 'pus_cell',
       'pus_cell_clumps', 'bacteria', 'blood_glucose_random', 'haemoglobin',
       'packed_cell_volume', 'white_blood_cell_count', 'red_blood_cell_count',
       'hypertension', 'diabetes_mellitus', 'coronary_artery_disease',
       'appetite'],
```

Figure 6. Index Result

Again, dataset was spitted into training & testing. These results are shared below. Our dataset has 15 features; those are directly applied to the machine learning models like Decision Tree, Support vector, KNN, Naive Bayes and their results are shown below. Figure 7 shows the Custom CNN (Sequential) models are created with the help of keras library.

```
df_model_eval = pd.DataFrame({'model': model_names, 'feature_count': feature_counts,
                              'acc': model_acc_scores,
                              'prc': model_prc_scores, 'rec': model_rec_scores,
                              'f1': model_f1_scores})
df_model_eval
```

	model	feature_count	acc	prc	rec	f1
1	dtc_model_1	15	0.991667	0.991667	0.991667	0.991667
2	nbc_model_1	15	0.950000	0.950000	0.950000	0.950000
3	svc_model_1	15	0.625000	0.625000	0.625000	0.625000
4	knc_model_1	15	0.708333	0.708333	0.708333	0.708333

Figure 7. Custom CNN (Sequential) models are created with the help of keras library

Table 1 shows the model 1 summary.

Table 1. Model 1: Summary

Model: "sequential"		
Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	1024
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 32)	2080
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 16)	528
dropout_2 (Dropout)	(None, 16)	0
dense_3 (Dense)	(None, 8)	136
dropout_3 (Dropout)	(None, 8)	0
dense_4 (Dense)	(None, 4)	36
dropout_4 (Dropout)	(None, 4)	0
dense_5 (Dense)	(None, 1)	5

Total params: 3,809
Trainable params: 3,809
Non-trainable params: 0

SGD optimizers are used in machine learning to find the best features for the training dataset by performing stochastic gradient descent. These are, in general, used to identify the optimal features of the training set that will be used in the subsequent prediction of new test data. The objective is to identify the most effective features for predicting the test data. The optimal number of features used in the prediction varies from one to multiple.

```
Epoch 11/20
9/9 [==============================] - 0s 2ms/step - loss: 0.0366 - accuracy: 0.9750
Epoch 12/20
9/9 [==============================] - 0s 2ms/step - loss: 0.0399 - accuracy: 0.9393
Epoch 13/20
9/9 [==============================] - 0s 3ms/step - loss: 0.0486 - accuracy: 0.9250
Epoch 14/20
9/9 [==============================] - 0s 3ms/step - loss: 0.0439 - accuracy: 0.9464
Epoch 15/20
9/9 [==============================] - 0s 3ms/step - loss: 0.0447 - accuracy: 0.9500
Epoch 16/20
9/9 [==============================] - 0s 2ms/step - loss: 0.0407 - accuracy: 0.9393
Epoch 17/20
9/9 [==============================] - 0s 2ms/step - loss: 0.0404 - accuracy: 0.9500
Epoch 18/20
9/9 [==============================] - 0s 3ms/step - loss: 0.0371 - accuracy: 0.9607
Epoch 19/20
9/9 [==============================] - 0s 3ms/step - loss: 0.0398 - accuracy: 0.9607
Epoch 20/20
9/9 [==============================] - 0s 3ms/step - loss: 0.0415 - accuracy: 0.9357
'========================'
Seq model 1 :  with SGD Optimizer Accuracy: 0.992
```

Figure 8. Accuracy

SGD optimizer is an optimisation technique used in machine learning models to find the best parameters or hyper parameters and hence it is a very powerful technique as shown in Figure 8. Figure 9 Shows the performance of model 1.

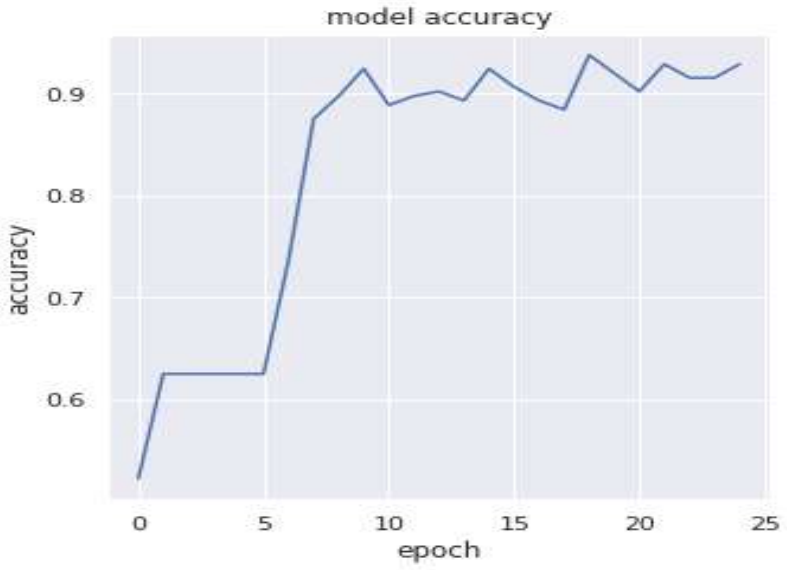

Figure 9. Performance level for Model 1

Table 2 shows the summary of the model 2.

Table 2. Model -2: Summary

Model: "sequential_1"		
Layer (type)	Output Shape	Param #
dense_6 (Dense)	(None, 50)	800
dropout_5 (Dropout)	(None, 50)	0
dense_7 (Dense)	(None, 30)	1530
dense_8 (Dense)	(None, 20)	620
dense_9 (Dense)	(None, 10)	210
dense_10 (Dense)	(None, 2)	22

Total params: 3,182
Trainable params: 3,182
Non-trainable params: 0

Sequential Model 2 with ADAM optimizer

```
Epoch 14/20
2/2 [==============================] - 0s 11ms/step - loss: 0.4202 - accuracy: 0.8929
Epoch 15/20
2/2 [==============================] - 0s 9ms/step - loss: 0.4069 - accuracy: 0.9179
Epoch 16/20
2/2 [==============================] - 0s 7ms/step - loss: 0.3786 - accuracy: 0.9071
Epoch 17/20
2/2 [==============================] - 0s 9ms/step - loss: 0.3702 - accuracy: 0.8857
Epoch 18/20
2/2 [==============================] - 0s 8ms/step - loss: 0.3455 - accuracy: 0.9107
Epoch 19/20
2/2 [==============================] - 0s 8ms/step - loss: 0.3197 - accuracy: 0.9250
Epoch 20/20
2/2 [==============================] - 0s 5ms/step - loss: 0.3067 - accuracy: 0.9286
'========================'
 Seq model 2: with Adam Optimizer Accuracy: 0.917
```

Figure 10. Adam Optimizer Accuracy

The Adam Optimizer optimizes a weight parameter for a neural network model. Adam optimizers are commonly used in machine learning and for optimizing gradient methods. The Adam optimizer uses a stochastic gradient descent algorithm to find the function that best fits the training data. It also uses a momentum term, to improve performance of the training algorithm as shown in Figure 10. Adam Optimizer is an algorithm that finds the best-fit model based on the least-squares objective function. It is used to automatically generate new models for a variety of classification tasks. Figure 11 shows the Model 2 Accuracy.

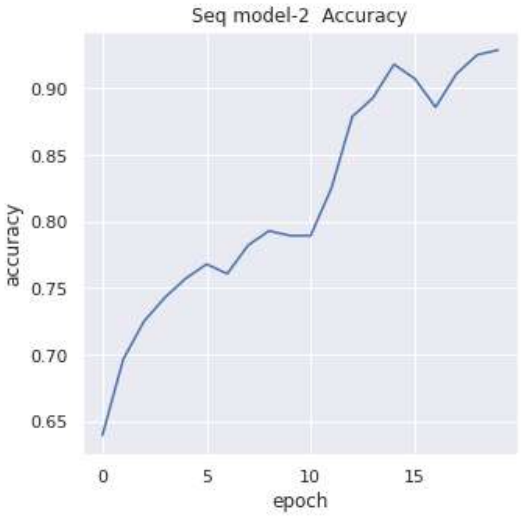

Figure 11. Model 2 Accuracy

CONCLUSION

In our work, by the help of RFE feature selection process to get the important features, applied to the classifier (both ML & DL) which is based on our set of machine learning & custom CNN models implementation. Even though machine learning models ("dtc_model") getting 99.16% accuracy, but our customized CNN model-1 with SGD optimizer getting 99.2% accuracy, getting slight better accuracy when compared with dtc_model. Therefore, CNN model-1 is best model for our CKD perdition approach**.** We will utilize data from CKD patients with 14 variables, as well as several machine learning approaches such as Decision Tree, SVM, and CNN model. To create an efficient machine learning model with the highest accuracy (by comparing several machine learning models) in predicting whether a person has CKD and, if so, how severe it is. Right now, we have less than 1000 records are taken up for this prediction. In future, we can gather huge volume of patient's data; it will process & applied towards our customized Deep CNN model (more features added).

REFERENCES

[1]. KA Padmanaban, and G Parthiban, 2016, "Applying machine learning techniques for predicting the risk of chronic kidney disease," *Indian J. of Science and Tech.*, **9(29)**, pp. 1-6.

[2]. P. Chittora, S. Chaurasia, P. Chakrabarti, G. Kumawat, T. Chakrabarti, Z. Leonowicz, M Jasiński, Ł Jasiński, R Gono, E Jasińska and V Bolshev, 2021, "Prediction of chronic kidney disease-a machine learning perspective," *IEEE Access*, **9**, pp.17312-17334.

[3]. J. Qin, L Chen, Y Liu, C Liu, C Feng and B Chen, 2019, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, **8**, pp. 20991-21002.

[4]. M. Almasoud, and T.E Ward, 2019, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," *Int. J. of Soft Computing and Its Applications*, **10(8)**.

[5]. MU Emon, R Islam, MS Keya and R Zannat, 2021, "Performance Analysis of Chronic Kidney Disease through Machine Learning Approaches," *In2021 6th Int. Conf. on Inventive Computation Technologies (ICICT)*, pp. 713-719.

[6]. AJ Aljaaf, D Al-Jumeily, HM Haglan, M Alloghani, T Baker, AJ Hussain, and J Mustafina, 2018, "Early prediction of chronic kidney disease using machine learning supported by predictive analytics," *In2018 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1-9.

[7]. A Subasi, E Alickovic and J Kevric, 2017, "Diagnosis of chronic kidney disease by using random forest," *InCMBEBIH 2017*, pp. 589-594.

[8]. H L Meghana, S Vaishnavi Kuber, B S Yamuna, T L Varshitha and Prof. B. M Vikrant, 2021, "Chronic Kidney Disease Prediction using Neural Network and ML Models" *In Int. J. of Eng. Res. & Tech. (IJERT)* **9(8)**, pp. 6-9.

[9]. K. M. Zubair Hasan, and Zahid Hasan, 2019, "Performance evaluation of ensemble-based machine learning techniques for prediction of chronic kidney disease." In *Emerging Res. in Computing, Information, Comm. And Applications*, pp. 415-426.

[10]. A Maurya, R Wable, R Shinde, S John, R Jadhav and R Dakshayani, "chronic kidney disease prediction and recommendation of suitable diet plan by using machine learning," *In2019 Int. Conf. on Nascent Technologies in Eng. (ICNTE)*, pp. 1-4.

[11]. J. P. T. M.Noordhuizen, and H. M. Metz, 2005, "Quality control on dairy farms with emphasis on public health, food safety, animal health and welfare," *Stočarstvo: Časopiszaunapređenjestočarstva*, ***59*(1)**, pp.39-55.

[12]. A. Pandey, and G. Prakash, 2019, "Deduplication with attribute-based encryption in E-health care systems," *Int. J. of MC Square Sci. Res.*, ***11*(4)**, pp.16-24.

[13]. S. Araya, B. Abera, and M. Giday, 2015, "Study of plants traditionally used in public and animal health management in SehartiSamre District, Southern Tigray, Ethiopia," *J. of ethnobiology and ethnomedicine*, ***11*(1)**, pp.1-25.

[14]. M. S. Ali-Shtayeh, R. M. Jamous, and R. M. Jamous, 2016, Traditional Arabic Palestinian ethno veterinary practices in animal health care: a field survey in the West Bank (Palestine). *J. of ethno pharmacology*, ***182***, pp.35-49.

[15]. M. M. Curran, G. Feseha, and D. G. Smith, 2005, "The impact of access to animal health services on donkey health and livelihoods in Ethiopia," *Tropical animal health and production*, ***37*(1)**, pp.47-65.

[16]. C. W. Young, V. R. Eidman, and J. K. Reneau, 1985, "Animal health and management and their impact on economic efficiency," *J. of Dairy Science*, ***68*(6)**, pp.1593-1602.

[17]. B Pattanaik, and S. Murugan, 2017, "Cascaded H-Bridge Seven Level Inverter using Carrier Phase Shifted PWM with Reduced DC sources." *Int. J. of MC Square Sci. Res.* **9(3)**, pp. 30-39.

[18]. A. Hussaindeen, S. Iqbal and T. D. Ambegoda, 2022, "Multi-Label Prototype Based Interpretable Machine Learning for Melanoma Detection," *Int. J. Adv. Sig. Img. Sci*, **8(1)**, pp. 40–53

[19]. F.J. Xavier, 2019, "Separation and Classification of Fetal ECG Signal by Enhanced Blind Source Separation Technique and Neural Network," *Int. J. Adv. Sig. Img. Sci*, **5(2)**, pp. 7–14.

[20]. Y. B. Bakare, and M. Kumarasamy, 2021, "Histopathological Image Analysis for Oral Cancer Classification by Support Vector Machine," *Int. J. Adv. Sig. Img. Sci*, **7(2)**, pp. 1–10.

[21]. E. Balamurugan, and J. Akpajaro, 2021, "Genetic Algorithm with Bagging for DNA Classification," *Int. J. Adv. Sig. Img. Sci*, **7(2)**, pp. 31–39.