A NEW DEEP LEARNING MODEL FOR CHRONIC KIDNEY DISEASE PREDICTION

Resul Özdemir
ORCID: 0000-0003-3678-812X
Murat Taşyürek
ORCID: 0000-0001-5623-8577
Veysel Aslantaş
ORCID: 0000-0002-0952-0315

ABSTRACT

Chronic Kidney Disease (CKD) is a life-threatening disease. Early diagnosis of mortal illnesses is of utmost importance. Utilizing historical data for the prediction of future health conditions is the easiest, less costly and non-invasive way. This approach yields to early detection of diseases as well. Machine Learning (ML) and Deep Learning (DL) models are the most common technics for intelligent estimations. In this paper, we proposed a new 1D-CNN deep learning model to predict CKD problems. The proposed model is compared with K- Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Ma- chine (SVM) and Random Forest (RF) algorithms. Hyperparameters affect the model performances of ML and DL algorithms. In this study, hyperparameter optimization is employed for all models using the Grid Search CV optimization algorithm. In order to improve model quality, 10-fold cross-validation is conducted for all ML models. The comparison metrics, accuracy, precision, recall and F1 score are utilized for performance evaluations. The accuracy metric is compared for both training and test data sets. The remaining metrics, precision, recall and F1 score, are calculated for only the test data set. The proposed 1D-CNN model achieved competitive results among all machine learning models. We have reported 98.93% accuracy for both training and test data sets. Moreover, 98.93% precision is calculated for the test data set. The recall metric value is 98.93% for the test data set. Finally, the F1 score is calculated as 98.93%, which is the best achieve- ment among all ML models in this study.

Keywords: chronic kidney disease, machine learning, deep learning, hyperparameter optimization

INTRODUCTION

Kidneys are the organs located on the lateral parts and toward the back of the human body. The main purpose of kidneys is to segregate toxins and water by filtering blood. In addition to the fact that the kidneys have an important role in the human body, they also cause vital illnesses. Chronic Kidney Disease (CKD) is a life-threatening issue worldwide. Approximately 14% of the world population suffers from this disease, and it brings about more deaths than prostate and breast cancer Ekanayake and Herath (2020).

Prediction of CKD is a life-saving action that leads to early detection of illness. There have been many measurements to classify CKD, such as age, blood pressure, amounts of sodium, potassium, haemoglobin, etc. Rubini et al. (2015). In order to create a relationship between these measurements and disease, Machine Learning (ML) and Deep Learning (DL) models are utilized.

In the literature, K-Nearest Neighbour (KNN), Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF) algorithms are popular ML algorithms. Convolutional Neural Networks (CNN) is a DL method which has better performance for two-dimensional and three-dimensional data, especially image data Arslan and Tasyurek (2022), Ozturk et al. (2023), Moustapha et al. (2022), Tasyurek and Ozturk (2022), Ulutas and Aslantas (2023), Akbulut and Aslantas (2023). On the other hand, the One-Dimensional Convolutional Neural Networks (1D-CNN) method is an effective DL technic for one-dimensional data sets. In this study, since the CKD data set is a tabular data set, we implemented a 1D-CNN deep learning model. The proposed DL method has competitive results among powerful ML algorithms.

RELATED WORKS

Islam et al. (2020) have studied CKD risk factor prediction using three different ML algorithms. In the study, ML algorithms Naïve Bayes (NB), Random Forest (RF), and Simple Logistic Regression (SLR) are compared.

The accuracy values are compared in this study. The NB, RF and SLR algorithms have achieved 93.91%, 98.89%, and 94.77% accuracy values, respectively. RF algorithm has the best performance in this study.
Arulanthu and Perumal (2019) have compared four different ML algorithms to predict CKD. Jrip, NB, Sequential Minimal Optimization (SMO) and K-Nearest Neighbor (KNN) algorithms were evaluated in this study. The obtained accuracy scores are 98.80%, 95.00%, 97.75% and 95.75%, respectively. The Jrip model has achieved the best performance among the four ML models.
Revathy et al. (2019) have compared machine learning models for CKD prediction. SVM, DT and RF models were employed for the prediction. RF algorithm has the best performance in terms of accuracy metrics among all algorithms.
Pasadana et al. (2019) have studied different decision tree models for CKD problems. Random forest, DecisionStump, HoeffdingTree, J48, CTC, J48graft, LMT, NBTree, RandomTree, REPTree, and SimpleCart models were utilized for the CKD data set. The random forest model achieved the best performance among tree-based models.
Lambert and Perumal (2021) performed Feature Selection (FS) for CKD prediction. Metaheuristic optimization techniques were utilized for the FS process. In the study, Ant Colony Optimization (ACO) algorithm, Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) algorithms were compared. The ACO-FS approach has produced the best solution for CKD prediction.
Nishanth and Thiruvaran (2017) have proposed an FS method based on KNN and Linear Discriminant Analysis (LDA). In the study, the obtained precision values are over 95% for the data set that emerged after the LDA-based feature selection process. LDA-based solutions have been reported to be much more successful than KNN-based solutions in this paper.
Apart from the estimation of CKD risk factors, studies suggest a suitable diet program for patients Maurya et al. (2019). In this study, three new features, which were Safe, Caution and Danger, were derived according to the amount of potassium in the data set. A system design for a diet program with newly released features is proposed.
In addition, Arulanthu and Perumal (2021) have another study that emphasizes the importance of CKD. This literature review study projects on the benefits of early prediction of CKD.
In this study, we have proposed a new 1D-CNN deep learning model. Moreover, the performances of KNN, DT, RF, SVM and the proposed 1D-CNN model are compared.

BASIC CONCEPT

In this section, the data used in this study is explained with all features and data types. Furthermore, the machine learning algorithms are explained to predict the chronic kidney disease classes.

Data

The features in the Chronic Kidney Disease dataset Rubini et al. (2015), which has been published in UCI Machine Learning Repository, are introduced. Table 1 lists the feature names, short descriptions of the relevant features, and the data type of each field in the dataset. The dataset contains 400 sample data and 25 features. There has been a Class feature, the target field for disease prediction. The remaining 24 features include 11 numerical and 14 categorical data types. In Table 1, the numerical features are listed with metrics. Likewise, the categorical features are shown with possible category options.

Table 1. Feature Description of Chronic Kidney Disease Data Set

Feature Name	Description	Data Type
Age	Age in years	Numeric
Blood Pressure	Amount of blood pressure in mm/Hg	Numeric
Specific Gravity	Specific gravity (1.005,1.010,1.015,1.020,1.025)	Categorical
Albumin	Albumin category (0,1,2,3,4,5)	Categorical
Sugar	Sugar category (0,1,2,3,4,5)	Categorical
Red Blood Cells	Red blood cells category (normal,abnormal)	Categorical
Pus Cell	Puss cell category (normal,abnormal)	Categorical
Pus Cell Clumps	Pus cell clumps category (present,notpresent)	Categorical
Bacteria	Bacteria category (present,notpresent)	Categorical

Blood Glucose Random	Amount of blood glucose random (mgs/dl)	Numeric
Blood Urea	Amount of blood urea (mgs/dl)	Numeric
Serum Creatinine	Amount of serum creatinine (mgs/dl)	Numeric
Sodium	Amount of sodium (mEq/L)	Numeric
Potassium	Amount of potassium (mEq/L)	Numeric
Hemoglobin	Amount of hemoglobin (gms)	Numeric
Packed Cell Volume	Amount of packed cell volume	Numeric
White Blood Cell Count	Count of white blood cells (cells/cumm)	Numeric
Red Blood Cell Count	Count of red blood cells (millions/cmm)	Numeric
Hypertension	Status of hypertension (yes,no)	Categorical
Diabetes Mellitus	Status of Diabetes Mellitus (yes,no)	Categorical
Coronary Artery Disease	Status of coronary artery disease (yes,no)	Categorical
Appetite	Status of appetite (good,poor)	Categorical
Pedal Edema	Status of pedal edema (yes,no))	Categorical
Anemia	Status of anemia (yes,no)	Categorical
Class	Target feature for classification (ckd,notckd)	Categorical

Methods

In this paper, the most popular ML algorithms, which are K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and 1D Convolutional Neural Network (1D-CNN), are introduced to classify CKD problem.

K-Nearest Neighbour (KNN)

KNN was proposed by Bay to solve classification problems of supervised learning Bay (1999). The basic concept of the algorithm is to find out the similarities between features and target classes. There is a parameter k that defines checking the number of the nearest data samples in the whole data set. The algorithm calculates the distances for all data samples in an iterative way and makes a decision about the target class among k neighbours.

Support Vector Machine (SVM)

SVM has been proposed by Cortes and Vapnik (1995) as a classification algorithm in the realm of machine learning. SVM calculates the margin size with hyper-planes that determine the classes in the given data set. The hyper-planes create a hidden layer that all the data set features are projected to this layer by a kernel function. This kernel function generates the boundary of the data set features to segregate the classes in the model.

Decision Tree (DT)

A decision tree is built to create a model for the classification operation in ML applications. The tree structure makes decisions with respect to features in the data set. The decisions are made from the root and leaves of the tree by calculation methods that reduce uncertainty. Loh (2011) has proposed the decision tree approach for classification and regression problems.

Random Forest (RF)

Random forest is a powerful ML algorithm originating from the decision tree approach. RF brings together decision trees to build a prediction model. All sub-models generate an estimation, and the RF algorithm ends up with the final decision according to the results of these sub-classifiers. Breiman (2001) proposed the RF algorithm.

1D Convolutional Neural Networks (1D-CNN)

Convolutional Neural Networks are deep learning methods to generate high-quality features of given data sets. Generally, CNN models are successful and popular methods in the realm of image-related operations. On the other hand, 1D-CNN models are the special forms of CNN models for one-dimensional data sets. 1D-CNN models generate deep features before neural network classification operations to stage better performance. There have been several studied regarding DL in the literature Arslan and Tasyurek (2022) Ozturk et al. (2023) Moustapha et al. (2022) Tasyurek and Ozturk (2022) Ulutas and Aslantas (2023) Akbulut and Aslantas (2023).

PROPOSED 1D-CNN MODEL

The architectural structure of the proposed 1D-CNN is illustrated in Figure 1. There have been six convolutional layers to obtain deeper features from inputs. The filter size of each convolutional layer is 16, 32, 16, 16, 8 and 4, respectively. There have been 24 features in the data set, and we determine the filter sizes with respect to the feature size accordingly. Rather than keeping the number of filters large and adding fewer layers, the number of filters is determined less and more layers are preferred in the proposed network design. The kernel size is set to 3 for all convolutional layers. The activation function of all convolutional layers is determined as Tanh function for this structure. All convolutional layers are followed by an activation layer which contains ReLU activation function. The Max Pooling layer is employed after convolutional layers. The pool size is set to 2 for this layer. Then, the flatten layer carries out the feature extraction for the neural network classifier. The neural network layer receives the inputs from the flatten layer. This layer contains 64 nodes with ReLU activation function. Finally, the proposed 1D-CNN network generates results with a SoftMax layer.

EXPERIMENTAL RESULTS

The data set is divided into two parts which are training and test data sets. The training set contains 70% of the original data set. On the other hand, the rest 30% of the data set is separated as the test set. In addition, a random 10% of the training set is allocated as the validation for hyperparameter optimization. In order to obtain the optimum parameter, set for all ML algorithms, the Grid Search CV algorithm is employed for hyperparameter optimization. Due to the fact that the data set contains a limited number of records, 10-fold cross-validation is utilized for better performance.

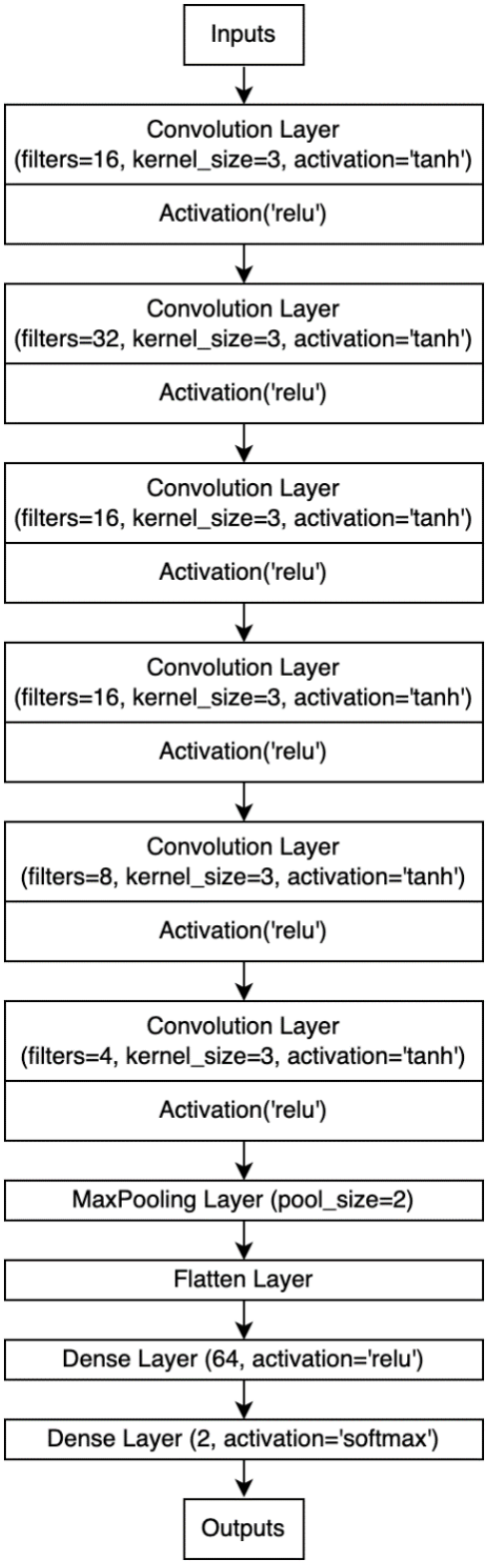


Figure 1 The proposed lD-CNN model architechture

Hyperparameter Optimization

The hyperparameters to be optimized for 1D-CNN, KNN, SVM, RF and DT algorithms are listed in Table 2. The parameters, optimizer, epoch and batchSize are optimized for the 1D-CNN model. The kernel, gamma and C parameters of the SVM algorithm are optimized. The parameters optimized for the RF algorithm are nEstimators, maxDepth, maxF eatures, minSamplesSplit and minSamplesLeaf. The parameters nNeighbors, metric and weights are optimized in the KNN algorithm. Finally, the maxDepth, minSamplesSplit, minSamplesLeaf and criterion parameters of the DT algorithm are optimized.

Table 2: ML Algorithms and Hyperparameters to be optimized

Algorithm	Hyperparameters
1D-CNN	optimizer, epoch, batchSize
SVM	kernel, gamma, C
KNN	nN eighbours, metric, weights
RF	nEstimators, maxDepth, maxF eatures, minSamplesSplit, minSamplesLeaf
DT	maxDepth, minSamplesSplit, minSamplesLeaf , criterion

The list of optimum hyperparameters obtained as a result of the optimization is listed in Table 3. According to results in this table, for 1D-CNN, optimizer is determined as Adam, epoch is 200 and batchSize is 100. The parameter kernel in the SVM algorithm is set as rbf. The gamma is set to 1 and C value is 2. The value of nNeighbors for the KNN algorithm is set to 4, metric value is minkowski, and weights parameter is determined as uniform. The value of nEstimators for the RF algorithm is 30, maxDepth is set to 10, maxFeatures is auto, minSamplesSplit is 2, and minSamplesLeaf is 1. Finally, for the DT algorithm, maxDepth is set to 10, minSamplesSplit is 6, minSamplesLeaf is 1, criterion is determined as entropy.

Table 3: ML Algorithms and Optimized Parameters with Values

Algorithm	Hyperparameters
1D-CNN	optimizer: **Adam**, epoch: **200**, batch Size: **100**
SVM	kernel: **rbf**, gamma: **1**, C: **2**
KNN	nNeighbours: **4**, metric: **minkowski**, weights: **uniform**
RF	nEstimators: **30**, maxDepth: **10**, maxFeatures: **auto**, minSamplesSplit: **2**, minSamplesLeaf: **1**
DT	maxDepth: **10**, minSamplesSplit: **6**, minSamplesLeaf: **1**, criterion: **entropy**

Comparison Metrics

Various comparison metrics are used according to the type of problem to examine the performance of ML and DL methods. In classification problems, accuracy, recall, precision, and F1 Score are commonly used metrics Ozturk et al. (2023).

These metrics are calculated as presented in Eq. 1.

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \\ Recall &= \frac{TP}{TP+FN} \\ Precision &= \frac{TP}{TP+FP} \\ F_1-score &= \frac{2\times Precision\times Recall}{Precision+Recall} \end{aligned} \quad (1)$$

In Eq. 1, when the metrics are analyzed respectively, accuracy shows the percentage of successfully classified samples, recall shows how many of the true positives were successfully predicted, precision refers how many positively predicted values are positive, and f1 score is a measure of the model's precision and robustness. In order to calculate these metrics, TP (true positive), TN (true negative), FP (false positive), and FN (false negative) values should be obtained. TP indicates the number of correctly predicted detections. Thus, this means kidney disease detection to anyone who has kidney disease. TN means a class cannot be detected when a class is not present. In other words, if a person does not have kidney disease, the disease is not detected here. FP is class detection when a class does not exist. Accordingly, when a person does not have the illness, this means kidney disease to that person. FN is that when it belongs to a class, that person's class cannot be detected. In other words, although the person has kidney disease, the disease cannot be detected.

Experiments

The results obtained in the experiments are listed in Table 4. The best results, according to the metrics, are highlighted in bold style. The highest accuracy value for the training dataset is 100%, obtained by the RF algorithm. The SVM algorithm has 99.29%, and the 1D-CNN algorithm has a 98.93% accuracy value. The accuracy values for the test data set are 98.33% for all the 1D-CNN, SVM and RF algorithms. KNN and DT algorithms have 97.50% and 96.67% accuracy values for the test data set, respectively. When we shed light on Precision values, the highest RF and SVM algorithms have reached 100% value. The highest recall value is 98.33% which belongs to the 1D-CNN model. Similarly, the F1 Score value reached the highest value of 98.33% in the 1D-CNN model.

Table 4: ML and DL Models Training and Test Results

Algorithm	Acc.(Training)%	Acc.(Test)%	Precision%	Recall%	F1 Score%
1D-CNN	98.93	**98.33**	98.33	**98.33**	**98.33**
SVM	99.29	**98.33**	**100**	95.8	97.87
RF	**100**	**98.33**	**100**	95.8	97.87
KNN	98.21	97.50	97.87	95.8	96.84
DT	99.29	96.67	95.83	95.83	95.83

Furthermore, the confusion matrix of the obtained results is also given for each method. Figure 2 belongs to the 1D-CNN model. Figure 3 is the confusion matrix of the SVM model. Figure 4 depicts RF confusion matrix results. Also, the confusion matrix of the KNN model and DT model are illustrated in Figure 5 and Figure 6.

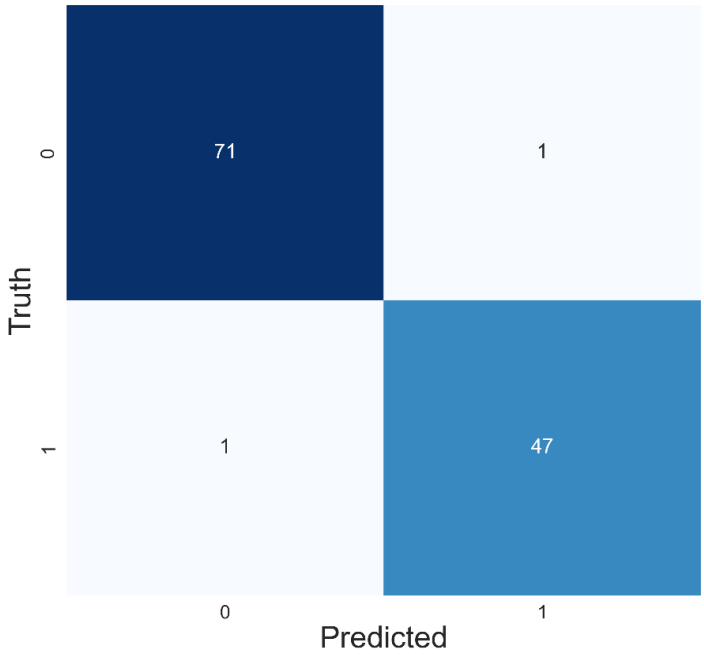


Figure 2 Confusion Matrix of 1D-CNN Model

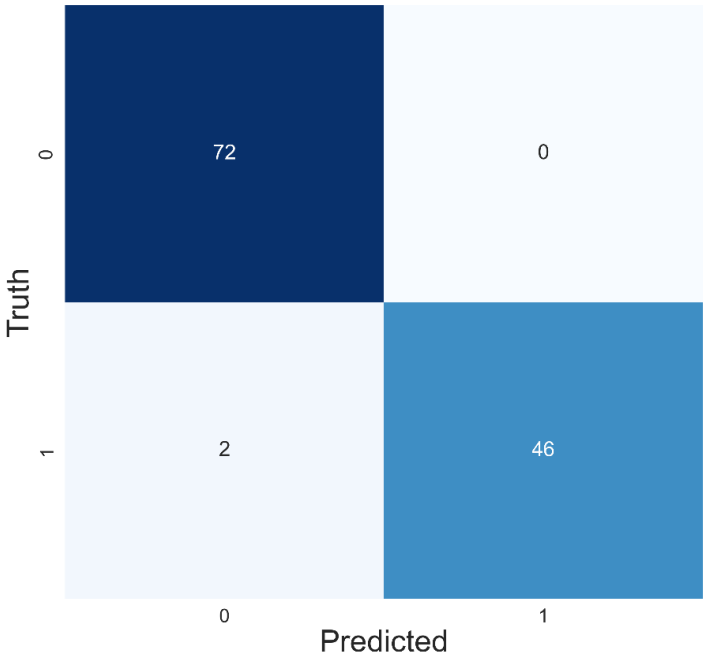


Figure 3 Confusion Matrix of SVM Model

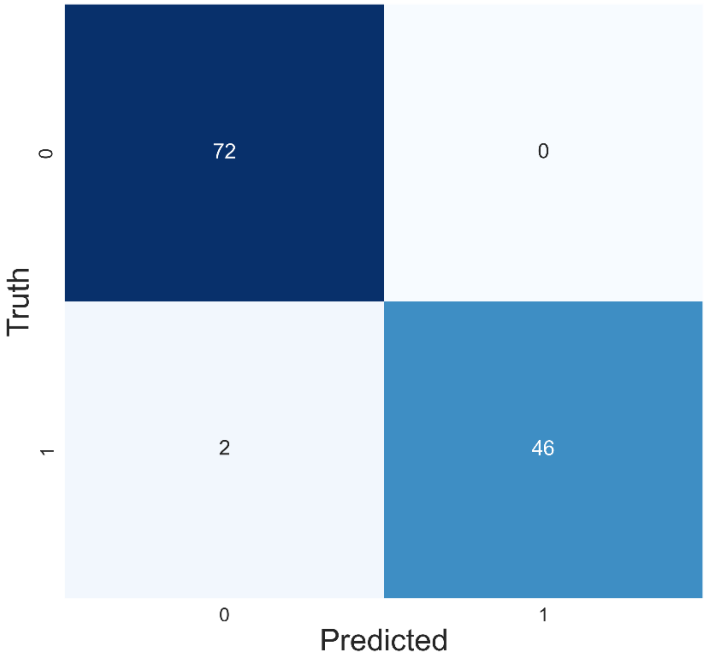


Figure 4 Confusion Matrix of RF Model

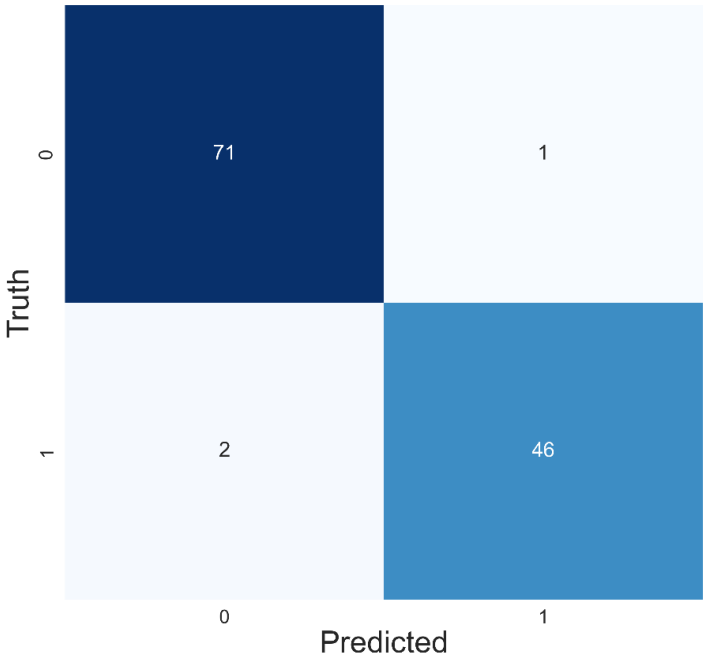


Figure 5 Confusion Matrix of KNN Model

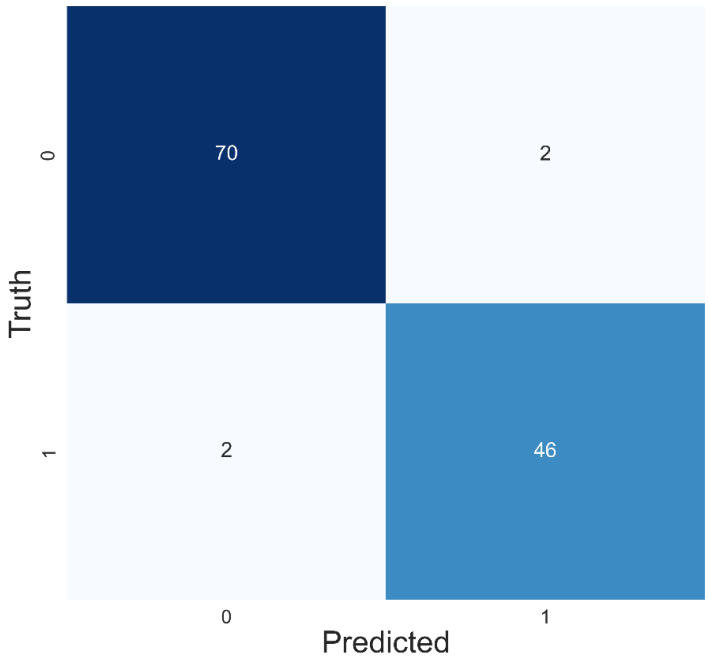


Figure 6 Confusion Matrix of DT Model

CONCLUSION

In this study, a new 1D-CNN deep learning model is proposed for the CKD prediction problem. Moreover, RF, SVM, DT and KNN machine learning methods are compared with the proposed DL method. At the first stage of this study, hyperparameters of all models are determined and optimized. The best parameter set of algorithms is employed for the same training and test data sets. Then, the models are built, and performance evaluation metrics values are obtained for performance comparison. The best accuracy performance for the training set belongs to the RF algorithm. On the other hand, the proposed 1D-CNN model, SVM and RF have the same accuracy value, which is the best performance. SVM and RF algorithms scored the best precision value. The proposed 1D-CNN model has the best scores in terms of recall and F1 score performance metrics. The proposed method produces competitive results with powerful ML algorithms in the literature.

REFERENCES

1. Akbulut, H. and Aslantaş, V. Evrişimli sinir ağı kullanarak çoklu-pozlamali görüntü birleştirme. *Gazi Universitesi Muhendislik Mimarlik Fakultesi Dergisi*, 38(3):1439–1452.
2. Arslan, R. S. and Tasyurek, M. (2022). Amd-cnn: Android malware detection via feature graph and convolutional neural networks. Concurrency and Computation: Practice and Experience, 34(23): e7180.
3. Arulanthu, P. and Perumal, E. (2019). Predicting the chronic kidney disease using various classifiers. In 2019 4th international conference on electrical, electron- ics, communication, computer technologies and optimization techniques (ICEEC- COT), pages 70–75. IEEE.
4. Arulanthu, P. and Perumal, E. (2021). Risk factor identification, classification and prediction summary of chronic kidney disease. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 14(8):2551–2562.
5. Bay, S. D. (1999). Nearest neighbor classification from multiple feature subsets. Intelligent data analysis, 3(3):191–209.
6. Breiman, L. (2001). Random forests. Machine learning, 45:5–32.
7. Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine learning, 20:273–297.

8. Ekanayake, I. U. and Herath, D. (2020). Chronic kidney disease prediction using machine learning methods. In 2020 Moratuwa Engineering Research Conference (MERCon), pages 260–265. IEEE.
9. Islam, M. A., Akter, S., Hossen, M. S., Keya, S. A., Tisha, S. A., and Hossain, S. (2020). Risk factor prediction of chronic kidney disease based on machine learn- ing algorithms. In 2020 3rd international conference on intelligent sustainable systems (ICISS), pages 952–957. IEEE.
10. Lambert, J. R. and Perumal, E. (2021). Optimal feature selection methods for chronic kidney disease classification using intelligent optimization algorithms. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 14(9):2886–2898.
11. Loh, W.-Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(1):14–23.
12. Maurya, A., Wable, R., Shinde, R., John, S., Jadhav, R., and Dakshayani, R. (2019). Chronic kidney disease prediction and recommendation of suitable diet plan by using machine learning. In 2019 International Conference on Nascent Technologies in Engineering (ICNTE), pages 1–4. IEEE.
13. Moustapha, M., Tasyurek, M., and Ozturk, C. (2022). A novel yolov5 deep learn- ing model for handwriting detection and recognition. International Journal on Artificial Intelligence Tools.
14. Nishanth, A. and Thiruvaran, T. (2017). Identifying important attributes for early detection of chronic kidney disease. IEEE reviews in biomedical engineering, 11:208–216.
15. Ozturk, C., Tasyurerek, M., and Turkdamar, M. U. (2023). Transfer learning and fine-tuned transfer learning methods' effectiveness analyse in the cnn-based deep learning models. *Concurrency and Computation: Practice and Experience*, 35(4): e7542.
16. Pasadana, I., Hartama, D., Zarlis, M., Sianipar, A., Munandar, A., Baeha, S., and Alam, A. (2019). Chronic kidney disease prediction by using different decision tree techniques. In Journal of Physics: Conference Series, volume 1255, page 012024. IOP Publishing.
17. Revathy, S., Bharathi, B., Jeyanthi, P., and Ramesh, M. (2019). Chronic kid- ney disease prediction using machine learning models. International Journal of Engineering and Advanced Technology, 9(1):6364–6367.
18. Rubini, L. J., Eswaran, P., and Soundarapandian, P. (2015). Chronic kidney disease data set. UCI Machine Learning Repository, CA, USA.
19. Tasyurek, M. and Ozturk, C. (2022). Ddl: A new deep learning based approach for multiple house numbers detection and clustering. Journal of the Faculty of Engineering and Architecture of Gazi University, 37(2).
20. Ulutas, H. and Aslantas, V. (2023). Design of efficient methods for the detec- tion of tomato leaf disease utilizing proposed ensemble cnn model. Electronics, 12(4):827.