Journal of Current Science and Technology, September-December 2022 Vol. 12 No. 3, 517-537

ISSN 2630 -0656(Online)

Cite this article: Pradeepa, P., & Jeyakumar, M. K. (2022, September). Data redundancy removal using K-MAD based self-tuning spectral clustering and CKD prediction using ML techniques. *Journal of Current Science and Technology*, *12*(3), 517-537. DOI:

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Data redundancy removal using K-MAD based self-tuning spectral clustering and CKD prediction using ML techniques

P. Pradeepa* and M. K. Jeyakumar

Department of Computer Applications, Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu 629180, India

*Corresponding author; E-mail: pradeepa20285@gmail.com

Received 8 April 2022; Revised 23 July 2022; Accepted 28 July 2022;
Published online 26 December 2022

Abstract

Chronic kidney disease (CKD) is one of the most complicated disorders, and it is found by gradual degradation of kidney function. People suffer to die several long-term complications like high blood pressure and heart and bone diseases. Hence, various automated early detection methods were developed to identify the disease at its early stage. Still, in numerous existing methods, the prediction level is inaccurate, so patients with low signs of CKD are found severe and undergo CKD treatments. This is because of the dataset's length and redundancy. To overcome these concerns, this paper focuses on increasing the prediction accuracy of CKD, utilizing an effective data mining approach. Therefore, to minimize the redundancy problem and high data dimension, this paper implemented the K-mad based self-tuning spectral clustering (KSSC) technique. The algorithm of self-tuning was designed to arrange data according to requirements and eliminate unnecessary data, resulting in a smaller data dimension. Various machine learning (ML) algorithms were used to verify the dimension-reduced data of Random Forest (RF), Artificial Neural Network (ANN), Deep Neural Network (DNN), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) classifier. Then the proposed technique was tested using various performance metrics in a Python environment, such as precision, f1_score, sensitivity, accuracy, specificity, and recall. The comparison study reveals that KNN and SVM deliver superior CKD predictions using a clustering method and attained 96% accuracy. Thus, the proposed KSSC shows essential information from healthcare centres and medical patient data, which is most helpful in assisting physicians in enhancing the accuracy of CKD diagnosis prior to a severe condition.

Keywords**:** *ANN; chronic kidney disease; DNN; KNN; machine learning algorithm; redundant self-tuning spectral clustering; SVM.*

1. Introduction

Chronic kidney disease is a disorder that occurs when a patient's kidney function deteriorates. Chronic kidney disease makes it harder to remove excess fluids from the body's bloodstream (Xiao et al., 2019). The last stage of kidney disease is renal disease, resulting in critically high amounts of fluid, electrolytes, and wastes in the body (Wang, Chakraborty, & Chakraborty, 2020). Weak bones, anaemia, nerve damage and high blood pressure are possible side effects. In order to avoid these concerns, the modern medical system developed various health monitoring systems to detect diseases at the initial

stage (Almasoud, & Ward, 2019). Therefore, health care providers and medical diagnosis systems generate a large quantity of data, such as images, video, text, and audio, that is both complex and voluminous and that cannot be processed and evaluated using traditional methods (Alloghani, Al-Jumeily, Hussain, Liatsis, & Aljaaf, 2020). To identify computer-assisted automatic disease, the machine learning approach delivers valid decision-making methodologies (Sobrinho et al., 2020). Machine learning (ML) is often used to improve the diagnostic process efficiency by automatically understanding current data and translating it into helpful information (Khan, Naseem, Muhammad, Abbas, & Kim, 2020). Machine learning is already being used to diagnose a variety of diseases, evaluate human health, and investigate disease-related.

Data mining and machine learning strategies provide the procedures and tools required to translate this heterogeneous information into actionable information (Bradley et al., 2019). Thus, the techniques are considered unsupervised learning and semi-supervised learning (Nusinovici et al., 2020). A semi-supervised technique defines a set of machine learning techniques that can train a model using both labelled and unlabeled data. The term "unsupervised" is often used interchangeably with "clustering" (Scholar, 2018). While the input instances are not class labelled, the learning technique can be performed unsupervised (Cheng, Huang, Zhang, Zhang, & Luo, 2021; Parmar et al., 2019b). The problem of detecting classes in data is solved using clustering algorithms (Shetty, Ahmed, & Naik, 2019). For dealing with statistical data analysis, clustering strategies are utilized. Clustering has a variety of uses in finance, health care, the internet, marketing research, computer science and a variety of other fields (Thongprayoon et al., 2021; Guo, Yu, Chen, & Zhao, 2020). Specifically, in medical applications, clustering can be used to identify a number of disorders (Parmar et al., 2019a; Wang, Ding, Wang, & Ding 2021).

Machine learning techniques such as Naive Bayes, decision tree, ANN, and SVM can be used to attain early predictions (Elhoseny, Shankar, & Uthayakumar, 2019; Hegde, & Mundada, 2020). The classifier can influence the dimension of the data when forecasting CKD (Lim et al., 2021). To address this difficulty, a clustering algorithm is utilized in the process of CKD prediction. Clustering is a method for discovering simar data based on their features. Researchers have introduced several different kinds of clustering techniques, such as Fuzzy C-means clustering, Density-based clustering, spectral based clustering, and K means clustering (Zelnik-manor, & Perona, 2004; Zhang, Li, & Yu. 2011; Alshammari, Stavrakakis, & Takatsuka 2021). This paper focuses on minimizing the burdens of CKD detection with high dimensional datasets. As a result, the proposed method includes a self-tuning spectral clustering algorithm based on redundancy. The main contribution to research is given as follows.

- An effective redundancy based Self-Tuning based Spectral Clustering Algorithm is presented for the early diagnosis of Chronic Kidney Disease.
- A K-mad based self-tuning spectral clustering (KSSC) to remove the redundant data.
- Using a dimension-reduced dataset, machine learning approaches are used to predict CKD.
- For effective early CKD detection, five different machine learning algorithms as, KNN, DNN, SVM, ANN, and RF, are utilized for classification.

The remaining part of the paper is organized as follows: section 2 will include a review of the article related to the CKD prediction model using clustering and machine learning techniques. Section 3 will include the background of the proposed methodology and the procedure of the proposed framework. Section 4 will include the result assembled through implementation. Finally, section 5 will conclude the entire research work.

1.1 Related Works

Several researchers have utilized dissimilar clustering techniques and machine learning methods for the prediction of CKD. Several articles are reviewed in this section.

Akben (2018) presented a new automatically diagnosis method for early-stage chronic kidney disease. The fundamental goal of this strategy was to help medical diagnosis based on developing the patient's and disease information, blood test and urine test. This research utilizes K-Means clustering algorithms in preprocessing stage. After preprocessing, there were three classification methods, Naive Bayes, KNN and SVM, were used to detect the CKD. These three classification techniques and clustering methods were utilized as the data mining approach. Clustering approaches were employed as a feature extractor as well as a

classifier. Lambert and Perumal (2021) used intelligent optimization techniques to develop an effective feature selection strategy for chronic renal disease classification. In order to reduce computational complexity and enhance classification performance in the diagnosis of CKD, this method developed a new methodology for CKD classification and prediction. In the CKD classification process, this method utilizes three different ways to feature selection such as Genetic Algorithm (GA), Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO). After feature selection, the extracted features were given to the Logistic regression (LR) classifier for diagnosis the CKD. Rady and Anwar (2019) presented the CKD prediction using data mining techniques. The early detection and diagnosis of CKD were thought to be critical in the disease's control and therapy. The large dimension of the dataset could make machine learning algorithms more difficult to train. Data mining techniques were developed to overcome these restrictions. For the detection of CKD, this methodology used four different machine learning strategies, Probabilistic Neural Network (PNN) algorithms, Multilayer Perceptron (MLP), Radial Bias Function (RBF), and SVM.

Zhang, Yang, Li and Fujita (2019) have solved various clustering problems using multiview and multitask based clustering. This research was designed with Laplacian Eigen maps as well as Locally Linear Embedding techniques. Initially converted, the samples into a general view space based on the various tasks. Afterwards, the samples are divided up into their own discriminative task space. Finally, use K-Means to cluster your data. To assess the clustering progress in future research, extracting complementary and shared features from many views was explored the clustering challenges. Aljarah, Mafarja, Heidari, Faris and Mirjalili (2020) presented a clustering strategy using Grey Wolf Optimization Tabu Search (GWOTS). To improve GWO's performance and to enable the search in the area of already proven optimal solutions, a TS-based technique was applied. The efficiency of GWOTS in comparison to earlier techniques was examined using thirteen clustering datasets based on statistical parameters. In addition to early convergence, in local optima, the GWO has conflict limits. The TS could be used in conjunction with the GWO to tackle these issues. The hybrid GWOTS could be utilized in a variety of clustering applications. Karthick (2017) suggested a procedure for nonlinear metric learning for machine learning systems. For the semi-supervised clustering, hierarchical forest clustering was used first, followed by the K-nearest neighbour algorithm for prediction. The performance of six datasets, including iris, glass, vowel, cancer, letter, and DNA, was assessed in terms of execution time and accuracy. Onan (2022) suggested a bidirectional convolution RNN architecture to split bi-directional LSTM and GRU layers for text sentiment classification. A group-wise enhancement technique was used for feature extraction through bidirectional layers. In addition, pooling layers and convolution layers extract high-level features as well as reduced feature space dimensions. Eleven sentiment classification benchmarks were utilized to analyze the predictive performance. This method was the first study in which 14 state-of-the-art architectures had been factually evaluated.

Onan (2019a) suggested a consensus clustering based-undersampling model for imbalanced learning. 44 small scale and 2 large scale imbalance benchmark datasets had been used for empirical analysis. In addition, five clustering algorithms and their combined models were used for consensus clustering schemes. Finally, in the classification stage, five supervised learning methods and three ensemble learner models were developed for prediction. Onan, Korukoğlu and Bulut (2016) presented predictive outcomes of five statistical keyword extraction methods on classification algorithms and ensemble methods for scientific text document classification. It was a comprehensive review comparing base learning algorithms with five ensemble learning models. The models were compared in terms of F-measure, accuracy and AUC. Two way ANOVA test was used to validate the empirical analysis. Onan (2019b) suggested a two-stage framework for topic extraction from the scientific literature. An improved word embedding scheme was developed to extract needed topics from text collections. An ensemble cluster model was introduced to enhance cluster performance. The empirical research demonstrates that the ensemble word embedding technique for topic extraction outperformed the baseline word vectors in terms of prediction ability.

Onan and Korukoğlu (2017) developed an ensemble approach for more robust and efficient feature subset selection. A genetic algorithm was

used to aggregate the individual feature lists. Outcomes show that the suggested aggregation model is an effective way for sentiment classification, outperforming individual filter-based feature selection methods. Onan, Korukoğlu and Bulut (2017) presented a hybrid ensemble pruning scheme based on clustering and a randomized search for text sentiment classification. To deal with the instability of clustering results, a consensus clustering scheme was created. The elitist Pareto-based multi-objective evolutionary method was used to explore the search space of candidate classifiers. This method was compared to three ensemble methods and three ensemble pruning algorithms in an experiment. Onan (2021a) suggested a deep learning-based approach to analyzing sentiment on product reviews gathered from Twitter. For sentiment analysis, the combined architecture of TF-IDF weighted Glove word embedding and CNN-LSTM architecture was utilized. The predictive effectiveness of several word embedding methods with various weighting functions was assessed in the empirical investigation.

Onan (2020) suggested a recurrent neural network (RNN) based model for opinion mining on instructor evaluation reviews. Using traditional machine learning techniques, ensemble learning methodologies, and deep learning architectures, researchers analyzed a corpus of 154,000 reviews. Three traditional text representation systems and four-word embedding schemes were used in the empirical study. Onan (2021b) suggested an efficient sentiment classification scheme with high predictive performance in MOOC reviews by pursuing the paradigms of ensemble learning and deep learning. The performance of traditional supervised learning methods, ensemble learning methods, and deep learning methods in terms of prediction was assessed. Machine learning, ensemble learning, and deep learning approaches were used to examine a corpus of 66,000 MOOC reviews for the evaluation assigned. Onan (2018a) suggested an extensive comparative analysis of different feature engineering schemes and five different base learners in conjunction with ensemble learning methods for text genre classification. An ensemble classification technique was developed based on the empirical analysis, which merges the Random Subspace ensemble of Random Forest with four types of characteristics.

Onan and Toçoğlu (2021) suggested an effective sarcasm identification framework for social media data by pursuing the paradigms of neural language models and deep neural networks. In this model, an inverse gravity moment based term weighted word embedding model with trigrams was developed to represent a text document. The suggested system was tested on a three-sarcasm identification corpus for the evaluation task. Three neural language models, two unsupervised term weighting functions, and eight supervised term weighting functions were tested in the empirical study. Onan (2019c) represented a deep learning-based approach to sarcasm identification. The predictive performance of a topic-enriched word embedding scheme was compared to that of traditional word-embedding schemes. Six subsets of Twitter messages, ranging from 5000 to 30.000, were considered in this model. Onan (2018b) developed an efficient multiple classifier approach to text categorization based on swarm-optimized topic modelling. Here, four different diversity metrics among ensemble classifiers were combined. A swarm intelligence-based clustering algorithm was used to partition the classifiers into a number of disjoint groups based on the combined diversity matrix, and one classifier from each cluster was chosen to build the final multiple classifier systems.

Artificial intelligence-based applications are essential in the development of healthcare industries. Generally, for the identification and classification of CKD, machine learning approaches are commonly used. The machine learning techniques contain two stages, training as well as a testing stage for classification and detection. Machine learning classifiers get complicated by high-dimensional datasets. To address the problems, a clustering approach for efficient prediction should be designed. The above-discussed methodologies used different clustering techniques like k-means clustering, partition-based clustering, and density-based clustering, but these have some restrictions, and they can be addressed in order to achieve optimum results. The proposed strategy was researched and applied to various machine learning algorithms to address the challenges of clustering techniques. The next section contains a full description of the proposed technique.

2. Objectives

This paper focuses on increasing the prediction accuracy of CKD, utilizing an effective data mining approach. Therefore, to minimize the redundancy problem and high data dimension, K-mad based self-tuning spectral clustering (KSSC) technique is proposed.

3. Proposed Methodology

CKD is one of the most important health concerns attributed to an increase in global occurrence and includes illnesses that gradually damage the kidneys and reduce the body's ability to conduct important tasks for a longer period of time. A K-mad based self-tuning spectral clustering (KSSC) is presented as an effective early detection of chronic kidney disease. Because of its increased prevalence worldwide, chronic kidney disease (CKD) is a major public health concern, causing the body to perform essential functions slowly and decreasing the ability to perform important functions, as well as starting to damage the kidneys once it cannot be detected initially. But, transplantation of a kidney is too costly; thus desired, an initial forecast of kidney failure. According to these reasons, many researchers developed various methods to discover early CKD disease using machine learning approaches. There have been many different clustering algorithms developed recently, and many of them produce low effective clusters. Complex data, more noise and unwanted data entities are the major issues present in multidimensional data. The clustering algorithms were used to deal with these issues, and different kinds of clustering based algorithms were utilized. The clustering procedure needs data reduction in order to achieve an effective execution time and minimize difficulties during clustering. Many different applications generally used dimension reduction strategies to overcome the problems of dimensionality.

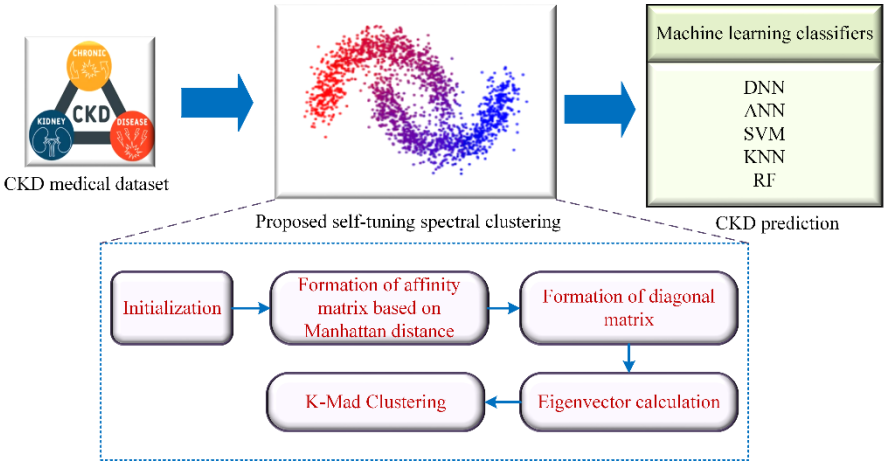

Figure 1 Proposed architecture of the KSSC algorithm for Chronic Kidney Disease Prediction

This proposed method introduced the KSSC to arrange the dimension as well as redundancy issues of high-frequency data in the CKD dataset. Furthermore, the clustering strategy reduces the difficulty of machine learning algorithms for CKD detection in terms of learning ability. The proposed KSSC is functional to the clusters according to desires and then eliminates the irrelevant data, reducing the dimension of the initial raw CKD data. Subsequently, for the purpose of predicting CKD, the dimension reduction data were fed into the classification procedure. The proposed method consists of three phases such as data collection, dimension reduction and classification. The initial process is the collection of the dataset from open source systems. After that, high-dimensional data was sent through a KSSC method, which minimizes the raw data's dimension. Finally, the ML techniques such as DNN, KNN, ANN, SVM, and RF are used to classify the dimension reduced data. Machine learning techniques' learning difficulties can be reduced by using the proposed dimension reduction data. The following section describes the step by step technique used in this proposed method.

3.1 Description of Dataset

The proposed method makes use of the CKD dataset from the machine learning repository UCI. The dataset was made up of 400 instances, and there were 250 people affected by CKD and 150 people who were not affected by CKD. Now, For CKD prediction, 24 features and two classifications were taken into account. The attributes were blood pressure, age, Hemoglobin, Red blood cells, anaemia, blood urea, blood glucose random, serum creatinine, specific gravity, pus cell clumps, pus cell, red blood cell count, hypertension, packed cell volume, white blood cell count, sodium, diabetes, sugar, potassium, pedal oedema, appetite, bacteria, coronary artery disease. CKD affected as well as not affected persons were the two classes. From that 400 instances, 224 were utilized for training the classier, and the rest of the data were utilized for testing the trained prediction algorithms. The performance of machine learning algorithms could be affected by large data dimensions. A proposed redundancy-based self-turning spectral technique was offered to improve performance and minimize data dimension. The following was the procedure for the proposed redundancy-based self-turning spectral technique.

3.2 Redundancy based Self-tuning Spectral Clustering Technique

After data collection, high dimensional data was given to the KSSC algorithm. The clustering strategy requires data reduction to give an efficient processing time and to decrease difficult clustering during clustering (Wen, 2020). Many researchers widely utilize dimension reduction techniques in various applications. Because of the data's multidimensional structure, as determined by using the Euclidean distance formula, which cannot effectively be characterized in high-dimensional space, the traditional clustering approach was easily trapped. To minimize and overcome these issues, the proposed method utilized KSSC for dimension reduction and data reduction. Algorithm 1 shows how the proposed KSSC algorithm works.

Algorithm 1: Pseudocode of KSSC technique

Input: Raw Data, Dimension, dataset, No. of cluster
Output: Clustering result
Step 1: Initialization
Step 2: Formation of Affinity matrix based on Manhattan Distance
Step 3: Formation of Diagonal Matrix
Step 4: Eigenvector Calculation
Step 5: K-Mad Clustering: Manhattan Distance based K-means clustering

In medical applications for CKD detection, high-dimensional data is particularly common. According to the redundant features and high dimensions, the affinity matrix could be corrupt. Therefore, in this study, to learn an appropriate affinity matrix, the affinity matrix learning technique was presented. The affinity matrix can be calculated using the K-means clustering. The affinity matrix was compared in this proposed redundancy-based self-tuning spectral clustering, and various enhancements were applied and explained below.

- A learning algorithm of affinity matrix can take into consideration sample local information, and important clustering information can be kept using this affinity matrix.
- The affinity matrix's most important sample relationship is saved since the number of neighbours for every sample is not fixed. As a result, it could be a more precise and simple method than KNN.
- The affinity matrix, as well as the crucial low-dimensional feature space, is adaptively updated by the self-tuning spectral clustering algorithm.

A KSSC using the affinity matrix removes the samples' inherent correlation as well as their redundant characteristics. Non-textual components in the CKD dataset, such as increased noise, question marks, and reduced clustering effectiveness. Pre-processing is important for reducing these problems since it removes characters non-textual from the stems and dataset. When dealing with high-dimensional data, calculating the affinity matrix might be difficult. The redundant self-tuning clustering algorithm was used, and original CKD medical data reduced in size. The redundancy based self-tuning clustering process is explained below.

Step 1: Initialization

The CKD dataset was initialized. Let us consider the dataset be

$S=\{s_n\}$, n=1,2,….N (1)

where X is the dataset, s_n is the nth sample of the dataset, N is the total samples in the CKD dataset. After initialization, the affinity matrix can be calculated. The affinity matrix is generated using a local scaling method.

Step 2: Formation of Affinity matrix based on Manhattan Distance

The affinity matrix can be useful for analyzing sample correlation in spectral clustering. Each sample can be represented as a vertex in a weighted graph with normal undirected edges, and the heat kernel function can be used to estimate the length of an edge that is important. In this step, the affinity matrix of the above dataset is formulated using the below equation (2).

$$A_{ij}=e^{\left(\frac{-d^2(s_i,s_j)}{\sigma^2}\right)} ;i\neq j \quad (2)$$

where, σ is the scale factor; the affinity matrix of the same data is zero that is denoted as $A_{ii}=0$. $d^2(s_i,s_j)$ is the distance function based on Manhattan distance which is given below

$$d^2(s_i,s_j)=\sum_{l=1}^{M}|s_{i,l}-s_{j,l}| \quad (3)$$

Where, $s_{i,l}$ is the l^{th} attribute of data sample s_i. *M* is the dimension of the data sample.

Step 3: Formation of Diagonal Matrix

The diagonal matrix *D* is formulated using affinity matrix (A_{ij}). The diagonal matrix is calculated using the equation given below

$$D_{ii}=\sum_{i,j=1}^{n} A_{ij} \quad (4)$$

Then the affinity matrix is normalized using the diagonal matrix as given below

$$L=D^{-1/2}(AD)^{-1/2} \quad (5)$$

Step 4: Eigenvector Calculation

In this step, the eigenvectors of the normalized affinity matrix were determined, which is represented below

$$X=\{x_c\} \text{ , } c=1,2,\ldots C \quad (6)$$

Where, 'x_c' is the c^{th} eigenvector of data sample X. *C* is the largest eigenvector.

Step5: K-Mad Clustering: Manhattan Distance based K-means clustering

In this step, the X is normalized to find the number of clusters using the equation given below

$$Y=X\Big/\sqrt{\sum_{c=1}^{C}(x_c)^2} \quad (7)$$

Each row of the '*Y*' is considered to be the centroid for clustering the data using the k-means algorithm. The set of data points is considered as X, let $X=\{x_1,x_2,x_3,\ldots..,x_n\}$ and the set of centres are denoted as $V=\{v_1,v_2,\ldots,v_c\}$.

Step 1: The cluster centres are represented as 'c' are chosen at random.
Step 2: The distance among each cluster centre and data points is estimated using Manhattan distance, which is given as equation (8).

$$Dist_{XY}=|X_{ik}-X_{jk}| \quad (8)$$

Step 3: Among the cluster centre and data points, the shortest distance is assigned for the entire cluster centres presented.
Step 4: The following expression (9) was used to find the new cluster centre,

$$V_i=\left(\frac{1}{ci}\right)\sum_{i}^{ci} x_i \quad (9)$$

Where 'ci' denotes the number of data points in ith cluster.
Step 5: Each data point's distance from the newly discovered clusters was again estimated.
Step 6: When the data points were not reassigned, then end the process, or steps from 3 to 5 were repeated.

The classes were categorized using the k-means clustering algorithm from the baseline CKD dataset. A machine learning approach for CKD prediction can be employed with the clustering dataset. Early CKD prediction can assist in decreasing the complication of learning systems by reducing the quantity of great dimensional data. The proposed redundancy based self-tuning clustering approach was analyzed through four different types of classifiers. A flowchart to detaily describe the flow of proposed redundancy is shown in fig 2.

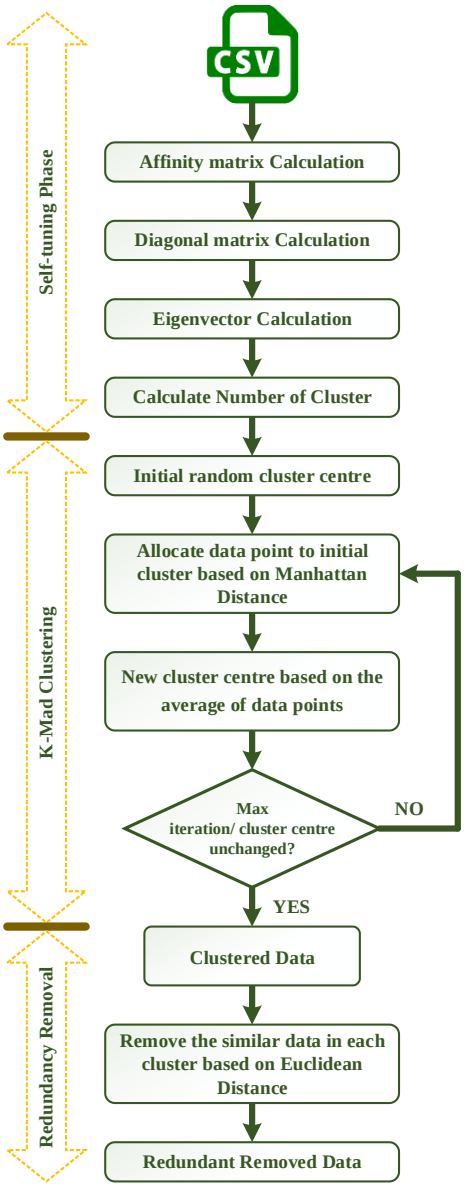

Figure 2 Flowchart for the proposed redundancy removal system.

3.3 Machine Learning Techniques Used for CKD Detection

This proposed CKD detection method considered four different machine learning classifications by using dimension reduced data. For high-dimensional analysis, the clustering procedures were utilized to decrease challenges in machine learning algorithms. In the medical industry, CKD diagnosis is a significant application because it allows doctors to identify a patient's condition before they suffer a serious illness. In order to examine CKD prediction, ML techniques were used. Mostly, due to the CKD redundancy and high dimension description, machine learning classifications are affected. Redundancy based self-tuning spectral clustering process was designed to group the needed data from the gathered CKD data to improve the performance of the ML techniques. By removing redundant data, the clustering method is beneficial for reducing the workload and improving the performance of the ML technique. The clustering technique was utilized to group classes based on the CKD data that had been collected. Machine learning approaches were used to detect CKD based on attributes using the clustering classes. The effectiveness of the

proposed KSSC was tested using an SVM and other classifiers such as DNN, ANN, RF and k-NN. To evaluate the accuracy of each classification method, the clustering data was fed into multiple classification algorithms. The utilized classification techniques are described below:

3.3.1. Modelling of DNN

The difference between DNN and Neural Network (NN) is that DNN has several hidden layers between the input and output layers. Deep Neural Networks (DNNs) had used in various fields like speech recognition, natural language processing, and computer version (Iliyas, Saidu, Dauda, & Tasiu, 2020). The DNN's operational methods are almost equal to the conventional neural network, and they differ from other neural networks based on the hidden layer. Hidden layers are several in between the input and the output layers. It can be capable of performing a huge amount of dataset and is highly effective compared to traditional machine learning techniques. In the DNN model, the input is sent through the hidden layer, and data can only move in one direction: From the input layer to the output layer, move forward. As a result, a deep neural network is referred to as a feedforward neural network. Another prominent feature of DNN is the avoidance of loops or cycles inside the network. Two phases are involved in DNN architecture such as training and testing (Lakshmanaprabu, Mohanty, Krishnamoorthy, Uthayakumar, & Shankar, 2019).

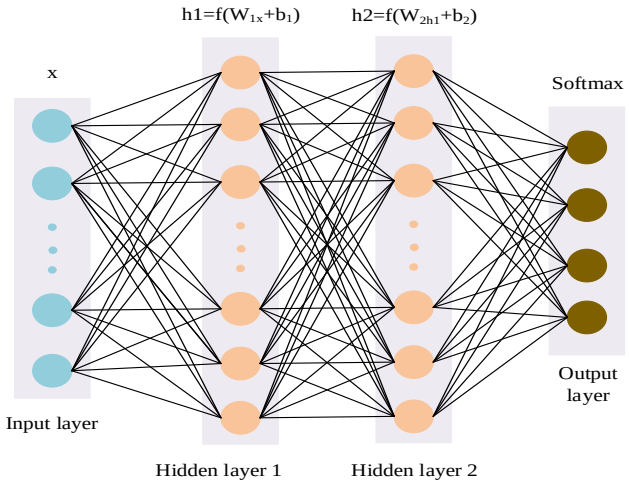

Figure 3 Schematic Diagram of DNN

The deep learning method is quite effective when a broader collection of samples is defined in the training phase. In the hidden layer, with the neuron bias, the weighted value of the input is subjected to the summing function, which is theoretically defined as the following equation (10).

$$C_{h(x)}=\left(\sum_{m=1}^{M} w_{xm}F_{I_j}\right)+b_x \quad (10)$$

Where, the interconnection weight amid the hidden layer and input is denoted as w_{xm} with M, K signifies the quantity of hidden and input neurons in the leading hidden layer and bias is represented as b_x. Constant value, $x = 1,2,\dots K$ number of hidden nodes as well as input. F_{I_j} denotes the chosen set of optimal features derived by redundancy-based clustering, which in turn represents the input of a deep learning algorithm that is available that ranges from $1 \leq m \leq M$ and $C_{h(x)}$ is the output of the entire hidden layer network. The activation function is the output of the hidden layer that is conveyed as exposed in equation (11)

$$a\left(C_{h(x)}\right)=\frac{1}{\left(1+e^{-C_{h(x)}}\right)} \quad (11)$$

The weight of connections between the hidden and output layers is represented by w_{xn}. The network's output is activated by the output layer activation function. In the network model, the error function is minimized and concludes the final output. It's measured using the mean square error (MSE). As shown in the equation, this is definitely

the difference between real and estimated probability values (12).

$$MSE=\frac{1}{T}\sum_{t=1}^{T}\left(C_{desired,T}-C_{estimated,T}\right)^2 \quad (12)$$

The term $C_{estimated,T}$ indicated the estimated and $C_{desired,T}$ is the outcome of the desired class. During the training phase, the error value decreases with each iteration. Once the network is trained, the system is tested using the trained net. The function of error reaches value is minimum, begin securing the sensitive data, and the network will cease iterating.

3.3.2. Modelling of ANN

ANN is commonly used to predict and/or classify any complex systems that are difficult to model using standard methods such as mathematical modelling. ANN model is used to detect CKD using the clustered data. For classification as well as recognition ANN method is widely applied for accurate prediction. There is no clear procedure for determining which ANN design and training progress delivers the best solution. The appropriate results in classification, as well as recognition, are obtained through trial and error. An artificial neuron is the primary element of the ANN structure, which is separated into three main parts. For instance, a weighted input is referred to as the input layer, the weight parameter is referred to as the hidden layer, and the summing part is referred to as the output layer (Almarashi, Alghamdi, & Mechai, 2018). The further value is transmitted to the membership function section of the second part of the neural network. The ANN's power is estimated based on the nonlinear function that is referred to as the membership function. Neurons in ANN were scheduled based on membership function to provide the optimum problem-solving capabilities. The training and testing phases of the ANN model have been processed successfully to predict the data (Senan et al., 2021). The clustered data was utilized for training the classifiers for acceptable error goals. After the ANN network had been trained, the weights and biases were fixed. After completing the training phase, the remaining clustering classes were utilized to test the trained ANN network for the prediction of CKD. The schematic diagram of ANN architecture is shown in figure 4.

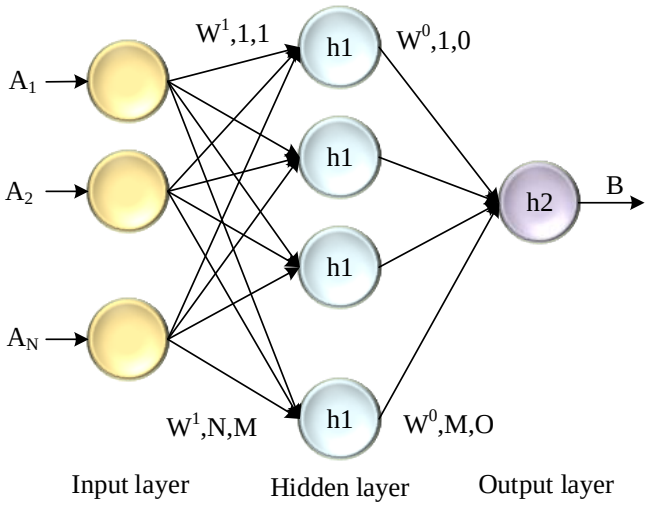

Figure 4 Architecture of the ANN model

Input layer: The input layer of the neural network is formed of multiple neurons. This input layer was in charge of gathering the clustered data from the KSSC algorithm. Then the input raw data sets were normalized to make the values within the limit values of the activation function. Within the range of activation functions, the input data were frequently normalized. The normalized outcomes are the better numerical accuracy standards of data in the input layer.

Hidden layer: The hidden layer is the intermediate layer between the input and output layers. With the

help of neurons, the hidden layer extracts corresponding values which are related to the detection system. Furthermore, if the membership function is available, it should be utilized in the hidden layer, and training data can be used to test the hidden node's weights.

Output layer: An active output layer is a final layer. Each neuron's previous layer outputs are added together to generate the overall training as well as testing results. The ANN structure was used to progress the training and testing phases. The performance evaluates the CKD prediction algorithm using the clustering technique.

3.3.3. Modelling of SVM

This method also uses the SVM classification approach for the diagnosis of CKD. SVM is computational learning, and Vapnik and collaborators were introduced to the SVM. The SVM classifier was created to help with nonlinear two-class classification problems. It minimizes risk and gives the greatest overall performance, and has been widely employed in a number of applications, including classification and recognition. The SVM is a supervised machine learning technique that uses a function to create output labels for new input values. It can reduce sample value error by boosting the model's generalization capacity and lowering structural risks (Ravindra, Sriraam, & Geetha, 2018).

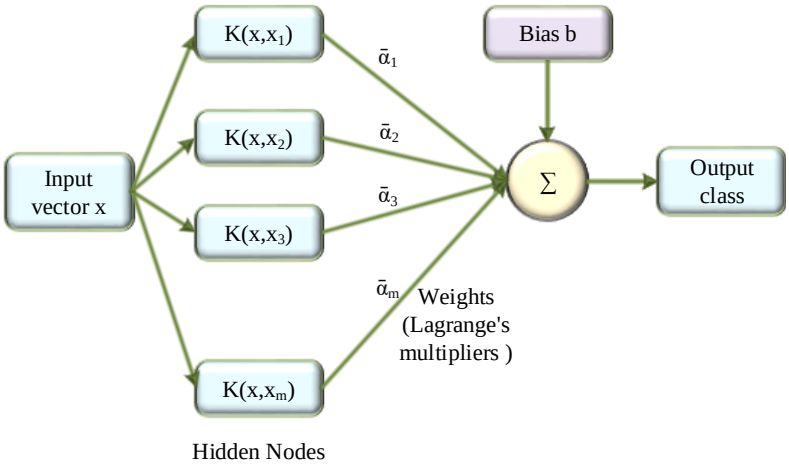

Figure 5 Schematic Diagram of SVM Architecture

The samples utilized in the classification consist of index values (I) and characteristics (S). Equation (13) defined a sample attribute-based dataset.

$$(X_a,Y_a),a=1,\ldots\ldots,S \text{ where } X_a \epsilon r^n,\ Y_a \epsilon \{-1,1\} \quad (13)$$

The hyperplane is expressed in equation (14),

$$(W,X)+P=0 \quad (14)$$

where, offset is denoted as *P*, an adjustable weight vector is represented as *W* is perpendicular to the hyperplane (Lakshmanaprabu et al., 2019). As illustrated in equation (15), the sample point distance is X_a to the classification hyperplane.

$$D_a=\frac{|WX_a+P|}{||W||} \quad (15)$$

The sample point's distance from the classification hyperplane maximization, which minimization is the same of $\frac{1}{2}||W||^2$. Constraint conditions, as given in the equation, are related to the function (16),

$$\begin{cases} \min \frac{1}{2}||W||^2 \\ \text{s.t. } Y|WX+P|-1\geq 0 \end{cases} \quad (16)$$

The Lagrangian function of a saddle point can be used to rule out the problem as the function is mentioned in equation (17).

$$\emptyset(W,B,\alpha_a)=\frac{1}{2}||W||^2-\sum_{a=1}^{s}\alpha_a[Y_a\,(WX_a+P)-1] \quad (17)$$

Equation (17) represents the Lagrangian function, $\alpha = [\alpha_1, \alpha_2, \ldots . \alpha_a], (\alpha_a \geq 0)$ defines the Lagrangian multiplier, which is used for dual programming purposes. In this way, better solutions for the weight factor as well as offset is achieved. The following equation (18) illustrates the final step of SVM, which is a decision-making function.

$$F(X)=sgn[\sum_{a=1}^{s} \alpha_a Y_a(XX_a)+P] \quad (18)$$

According to the SVM decision-making process, at each position within the specified space, the CKD prediction is categorized into two unique classes.

3.3.4. Modelling of KNN Classifier

KNN is one of the widely used classification algorithms that can be applied to reduce various data mining problems. An important technique of KNN employs a selection of classification in many applications relevant to its dataset by computing comparison directories, also known as distance functions. In the present research, KNN was applied to detect CKD using clustered data. Diagnoses of CKD were performed based on the similarity indices in KNN, with the best case being one with the most similarity [28]. The best quantity K values in KNN were associated with the measures used to predict CKD. In the KNN, to discover similarity functions in continuous and discrete variables in different functions were employed. For the estimation of discrete variable similarity functions, the hamming distance is commonly utilized. The computation of the continuous variable similarity function was performed through Euclidean distance. Other functions like Pearson correlation coefficients and Spearman correlation were also utilized in datasets. For selecting a dataset that differs from dataset to dataset according to the k value that was more significant in the KNN. The square root of the number of samples equals the k value, according to the empirical rule of thumb, which makes parameter tweaking difficult for various applications. In this method, Euclidean distance was utilized to estimate the similarity functions between the clustered datasets given in the following expression (19).

$$ED=\sqrt{\sum_{a=1}^{s} W_a(X_a-Y_a)^2} \quad (19)$$

Where, $\sum_{a=1}^{s} W_a = 1$ and $0<W_a<1$. The nearest neighbour classifier calculates the weighted Euclidean distance. Using the given equation, calculate the weighted Euclidean distance between two n-dimensional vectors. Clustered datasets were used to predict CKD using the distance function. The clusters were made through the proposed redundancy based self-tuned clustering approach, and the chosen clusters were fed into machine learning classification techniques, which yield better prediction findings. When used without the clustering process, machine learning approaches struggle to deal with redundancy in high-dimensional data while achieving accuracy across the board. The redundancy based self-tuning spectral clustering approach was proposed to simplify the difficulties of machine learning and improve the accuracy level of prediction in employed approaches. The primary aim is to predict CKD using the redundancy based clustered dataset, which was checked using five classification techniques, ANN, DNN, RF, KNN and SVM. The following part examines the simulation results obtained for the proposed redundancy based clustering algorithm with machine learning approaches.

4. Results and Discussion

4.1 Simulation Results

The KSSC with an ML technique was proposed for CKD prediction, which was implemented on python 3.8 environments, and the experimental examination has taken several parameters to reveal the performance of the proposed method. The proposed system focused on designing an effective CKD detection model for assisting physicians without the requirement of any numerical calculations. The initial process of the proposed method was gathering patient information from the UCI machine library. The next stage performed the clustering process utilizing the redundancy based self-tuning clustering algorithm. Then, the clustered dataset was provided to the machine learning approaches like SVM, ANN, RF, KNN and DNN for the diagnosis of CKD. For the classification process, initially, 80% of the clustered data were given to the machine learning techniques individually for training the classifiers. After completing the training process, the remaining 20% of the clustered data were given to the classifiers to test the trained classifiers. The performance of the

proposed method was computed based on considering two conditions, such as without clustering and with clustering, through five distinct machine learning approaches. Several performance metrics considered for analyzing the proposed models were sensitivity, f1_score, accuracy, recall, specificity and precision. Table 1 shows the execution parameters considered for the proposed KSSC based classification of CKD.

Table 1 Execution parameters considered for the proposed method

S.No	Methods	Parameter	Value
1	Dataset	Total attributes	25
2		Total samples	400
3		Training testing ratio	8:2
4	KNN	Distance formula	Euclidean
5		K value	3
6	ANN	Network	feedforward backpropagation
7		Learning rate	0.05
8		Activation function	Tansig function
9		output layer neurons	1
11		Input layer Neurons	2
12		hidden layer Neurons	20
13		Maximum epoch	500
14	SVM	Weight vector	0.725
15		Kernel	Gaussian
16		Kernel scale	1.2

The implementation outcome of the proposed method was collected according to the above-illustrated parameters. The following section contains a detailed overview of the proposed implementation outcomes.

4.2 Performance Analysis

The performance analysis of the proposed early CKD detection model carried several performance metrics. In the proposed method, the performance was calculated by utilizing statistical measurements. By using these attained statistical values, the performance of any system can be analyzed. The metrics considered for this analysis were accuracy, sensitivity, recall, specificity, precision and f1-score. False Positive (FP), False Negative (FN), True Negative (TN) and True Positive (TP) are statistical parameters that were used to estimate the statistical measurement. The statistical parameters explanation is delivered as below.

True positive (TP): The number of times an individual was identified as having CKD.

False Positive (FP): The number of times a person has been mistakenly identified as having CKD.

True Negative (TN): The number of instances of CKD was accurately identified as unaffected.

False Negative (FN): The number of times CKD was mistakenly labelled as unaffected.

The statistical measurement of the proposed KSSC with classifiers is shown in table 2. There were two conditions considered, such as with and without clustering, for the estimation of machine learning algorithms.

Table 2 Confusion matrix

Methods	With cluster				Without cluster			
	TP	TP	TP	TP	TP	FP	FN	TN
SVM	20	1	1	1	1	1	0	59
KNN	20	1	1	1	1	1	0	59
ANN	0	0	0	0	0	21	2	57
RF	15	0	0	0	0	1	15	49
DNN	25	0	0	0	0	2	22	31

The statistical measurements of the proposed method are computed based on considering with and without clustering through sensitivity, f1_score, accuracy, recall, specificity and precision. Table 3 illustrates values attained for the machine learning techniques with the clustering technique.

Table 3 Comparison of Machine Learning Methods with Clustering Technique

Techniques	Specificity	Accuracy	Precision	Recall	Sensitivity	F1_score
DNN	98%	98%	70%	90%	95%	85%
ANN	68%	70%	50%	70%	70%	65%
SVM	90%	85%	65%	83%	88%	80%
KNN	77%	75%	55%	75%	75%	70%
RF	60%	60%	47%	65%	65%	58%

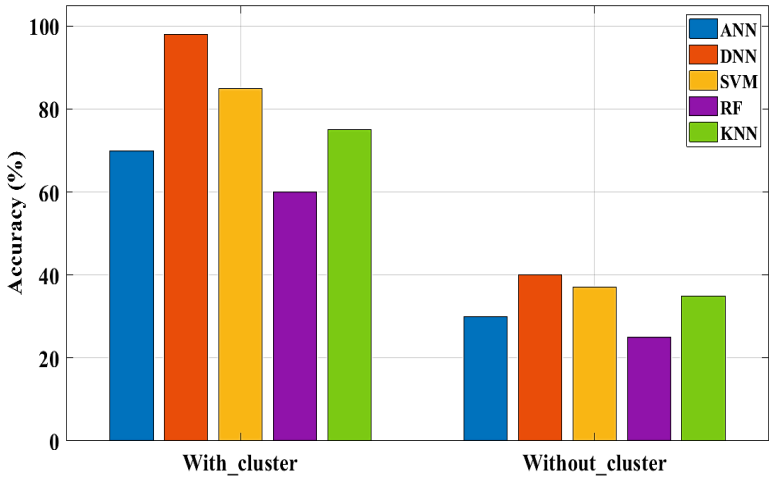

Figure 6 Comparison Analysis of Accuracy

Figure 6 illustrates the comparison analysis of accuracy among the utilized classification techniques based on with and without clustering process. Machine learning techniques include five methods: RF, SVM, ANN, \KNN and DNN are examined in two situations with and without clustering. To reduce complications of high dimensional data analysis using machine learning strategies, KSSC is proposed. After reducing the dimension of the dataset, to examine machine learning approaches behave, the clustered dataset was used in this study. The accuracy of the ANN classifier attained 70% with the clustering technique and 30% without the clustering technique. Meanwhile, the DNN classifier attained 98% of accuracy with the clustering technique and 40% of accuracy without the clustering technique. On the other hand, SVM classifier achieved 85% of accuracy with the clustering technique and 38% of accuracy without the clustering technique. The RF classifier achieved 60% of accuracy with the clustering technique and 25% of accuracy without the clustering technique. The KNN classifier obtained 75% of accuracy with the clustering technique and 35% of accuracy without the clustering technique.

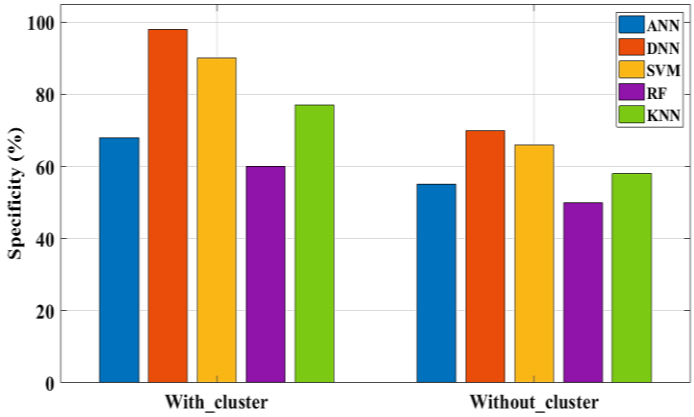

Figure 7 Comparison of Specificity

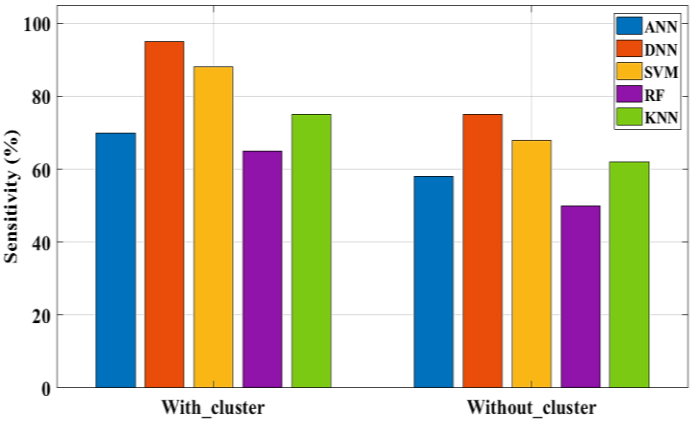

Figure 8 Comparison of Sensitivity

Table 4 Sensitivity comparison

Technique	With clustering	Without clustering
ANN	70	58
DNN	95	75
SVM	88	68
RF	65	53
KNN	75	62

The sensitivity with and without redundancy removal is given in table 4 and the specificity value attained for the five classifiers such as SVM, ANN, RF, KNN, and DNN without and with clustering techniques applied in the proposed method is illustrated in Figure 7. The specificity of the ANN classifier attained 70% with the clustering technique and 58% without the clustering technique. The DNN classifier attained 95% of specificity with the clustering technique and 75% of specificity without the clustering technique. The SVM classifier achieved 88% of specificity with the clustering technique and 68% of specificity without the clustering technique. The RF classifier achieved 65% of specificity with the clustering technique and 53% of specificity without the clustering technique. The KNN classifier obtained 75% of specificity with the clustering technique and 62% of specificity without the clustering technique. Figure 8 illustrates the sensitivity value achieved for the five distinct ML algorithms, such as KNN, SVM, DNN, RF and ANN, with and without clustering techniques applied in the proposed method. The sensitivity of the ANN classifier attained 70% with the clustering technique and 58% without the clustering

technique. The DNN classifier attained 93% of sensitivity with the clustering technique and 75% of sensitivity without the clustering technique. The SVM classifier achieved 89% of sensitivity with the clustering technique and 68% of sensitivity without the clustering technique. The RF classifier achieved 65% of sensitivity with the clustering technique and 50% of sensitivity without the clustering technique. The KNN classifier obtained 75% of sensitivity with the clustering technique and 62% of sensitivity without the clustering technique.

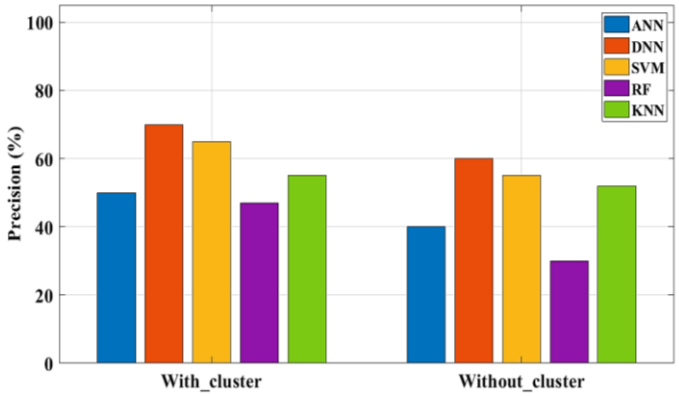

Figure 9 Comparison Analysis of Precision

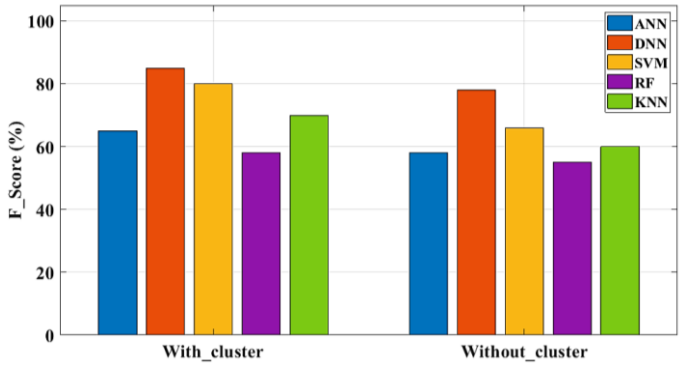

Figure 10 Comparison Analysis of F1_Score

Figure 9 illustrates the precision value attained for the five distinct ML algorithms like SVM, RF, KNN, DNN and ANN with and without clustering techniques applied in the proposed method. The precision of the ANN classifier attained 50% with the clustering technique and 40% without the clustering technique. The DNN classifier attained 70% of precision with the clustering technique and 60% of precision without the clustering technique. The SVM classifier achieved 65% of precision with the clustering technique and 56% of precision without the clustering technique. The RF classifier achieved 47% of precision with the clustering technique and 30% of precision without the clustering technique. The KNN classifier obtained 56% of precision with the clustering technique and 53% of precision without the clustering technique.

Figure 10 illustrates the f1_score value attained for the five distinct machine learning algorithms such as DNN, ANN, SVM, RF and KNN, with and without clustering techniques applied in the proposed method. The f1_score of the ANN classifier attained 63% with the clustering technique and 59% without the clustering technique. The DNN classifier attained 83% of f1_score with clustering technique and 79% of f1_score without clustering technique. The SVM classifier achieved 80% of f1_score with clustering technique and 68% of f1_score without clustering technique. The RF classifier achieved 58% of f1_score with clustering technique and 56% of f1_score without clustering technique. The KNN classifier obtained 65% of f1_score with clustering technique and 60% of f1_score without clustering technique.

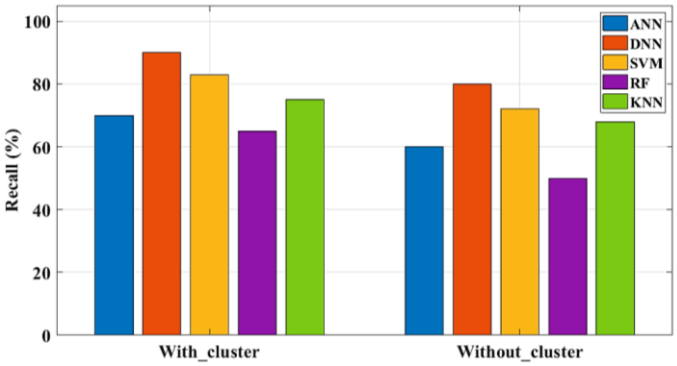

Figure 11 Comparison Analysis of Recall

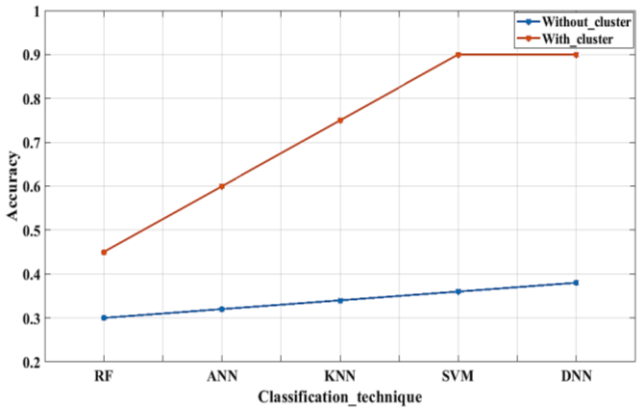

Figure 12 Overall Analysis of Accuracy in Machine Learning Techniques

Figure 11 illustrates the recall value attained for the five distinct machine learning algorithms such as ANN, KNN, SVM, DNN and RF, with and without clustering techniques applied in the proposed method. The recall of the ANN classifier attained 65% with the clustering technique and 60% without the clustering technique. The DNN classifier attained 90% of recall with the clustering technique and 80% of recall without the clustering technique. The SVM classifier achieved 81% of recall with the clustering technique and 72% of recall without the clustering technique. The RF classifier achieved 63% of recall with the clustering technique and 50% of recall without the clustering technique. The KNN classifier obtained 77% of recall with the clustering technique and 70% of recall without the clustering technique. Figure 12 illustrates the overall accuracy analysis with and without the clustering process. The accuracy value attained for the SVM, ANN, DNN, KNN and RF is 36%, 32%, 38%, 34%, and 30% without processing the clustering. Likewise, the accuracy values achieved by the five ML techniques, such as KNN, ANN, SVM, DNN, and RF are 90%, 60%, 90%, 75%, and 45% with the clustering technique. Finally, it determined that SVM and DNN with clustering produced the most accurate CKD prediction outcomes. In addition to the above performances, the clustering performance of the proposed techniques was compared in terms of computational complexity, Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), which is given in the table 5.

Table 5 Clustering Performance Comparision

Technique	Compution Complexity (in sec)	ARI	NMI
KSSC	73	0.68	0.71
Spectral Clustering	86	0.61	0.66
FCM	108	0.52	0.59
K Means	91	0.49	0.53

The clustering performance shown in table 5, compares the proposed KSSC with the conventional Spectral clustering, FCM and K-Means. The comparison was made in terms of computation complexity, it measures the execution time of the various clustering techniques. The execution time of the proposed KSSC is 73seconds, whereas other techniques like spectral clustering, FCM, and K-Means consume 86, 108 and 91 seconds, respectively. On the other hand, the proposed technique was also compared in terms of ARI and NMI. The proposed KSSC provides better performance in all metrics, proving that the proposed KSSC is better for clustering and redundancy removal.

5. Conclusion

In this research, an effective redundancy based self-tuning spectral clustering algorithm was proposed to minimize the complexities of machine learning techniques. Initially, real-time data and a benchmark of 400 samples with 25 attributes from the UCI machine learning repository were used to create a CKD patient. Then the gathered data were given to the KSSC, which decreases the dimension of redundancy and dataset. Then the machine learning algorithms were built using the redundancy removed data. Here five different machine learning techniques such as RF, KNN, SVM, DNN, and ANN were utilized to analyze the effectiveness of the redundancy removed data. RF These five different machine learning techniques were initially trained with 80% of clustered data (320 samples), and after training, the classifiers remaining 20% of clustered data were subjected to testing the trained model. Several performance measures were used to evaluate the proposed clustering algorithm's performance, such as recall, specificity, sensitivity, accuracy, f1_score and precision. In machine learning techniques, the proposed technique values of clustering are 98%, 85%, 75%, 70%, and 60%. Specificity values are 98%, 90%, 77%, 68%, and 60%. Sensitivity values are 95%, 88%, 75%, 70%, and 65%. Precision values are 70%, 65%, 55%, 50%, and 47%. F1_score values are 85%, 80%, 70%, 65%, and 58%. Recall values are 90%, 83%, 75%, 70% and 65%. According to this comparison analysis, DNN and SVM provide better CKD diagnosis with a clustering algorithm. Finally, it proves that clustering with machine learning techniques gives better specificity, f1_score, accuracy, sensitivity, recall and precision values. The proposed KSSC techniques for redundancy removal performed well with better performances. However, it has better performance, the integration of self-tuning with the K-Mad clustering performance is not justified for the nonlinear dynamic data. Moreover, the redundancy removal is not only enough to improve the prediction accuracy. So, in future, a data preprocessing technique with more features can be developed.

6. Acknowledgements

The authors would like to thank their university for allowing them to do this research work. There is no funding received.

7. Competing interests

The authors have no relevant financial or non-financial interests to disclose.

8. References

Akben, S. B. (2018). Early-stage chronic kidney disease diagnosis by applying data mining methods to urinalysis, blood analysis and disease history. *IRBM*, *39*(5), 353-358. doi.org/10.1016/j.irbm.2018.09.004

Aljarah, I., Mafarja, M., Heidari, A. A., Faris, H., & Mirjalili, S. (2020). Clustering analysis using a novel locality-informed grey wolf-inspired clustering approach. *Knowledge and Information Systems*, *62*(2), 507-539. doi.org/10.1007/s10115-019-01358

Alloghani, M., Al-Jumeily, D., Hussain, A., Liatsis, P., & Aljaaf, A. J. (2020). Performance-based prediction of chronic kidney disease using machine learning for high-risk cardiovascular disease patients. In *Nature-inspired computation in data mining and machine learning* (pp. 187-206). Springer, Cham. doi.org/10.1007/978-3-030-28553-1_9

Almarashi, A., Alghamdi, M., & Mechai, I. (2018). A new mathematical model for diagnosing chronic diseases (kidney failure) using ANN. *Cogent Mathematics & Statistics*, *5*(1), 1559457. doi.org/10.1080/23311835.2018.1559457

Almasoud, M., & Ward, T. E. (2019). Detection of chronic kidney disease using machine learning algorithms with least number of predictors. *International Journal of Soft*

Computing and Its Applications, *10*(8). DOI: 10.14569/IJACSA.2019.0100813

Almustafa, K. M. (2021). Prediction of chronic kidney disease using different classification algorithms. *Informatics in Medicine Unlocked*, *24*, 100631. https://doi.org/10.1016/j.imu.2021.100631

Alshammari, M., Stavrakakis, J., & Takatsuka, M. (2021). Refining a k-nearest neighbor graph for a computationally efficient spectral clustering. *Pattern Recognition*, *114*, 107869. https://doi.org/10.1016/j.patcog.2021.107869

Bradley, R., Tagkopoulos, I., Kim, M., Kokkinos, Y., Panagiotakos, T., Kennedy, J., ... & Elliott, J. (2019). Predicting early risk of chronic kidney disease in cats using routine clinical laboratory tests and machine learning. *Journal of veterinary internal medicine*, *33*(6), 2644-2656. https://doi.org/10.1111/jvim.15623

Cheng, D., Huang, J., Zhang, S., Zhang, X., & Luo, X. (2021). A novel approximate spectral clustering algorithm with dense cores and density peaks. *IEEE transactions on systems, man, and cybernetics: systems*, *52*(4), 2348-2360. DOI: 10.1109/TSMC.2021.3049490

Elhoseny, M., Shankar, K., & Uthayakumar, J. (2019). Intelligent diagnostic prediction and classification system for chronic kidney disease. *Scientific reports*, *9*(1), 1-14. doi.org/10.1038/s41598-019-46074-2

Guo, Y., Yu, H., Chen, D., & Zhao, Y. Y. (2020). Machine learning distilled metabolite biomarkers for early stage renal injury. *Metabolomics*, *16*(1), 1-10. https://doi.org/10.1007/s11306-019-1624-0

Hegde, S., & Mundada, M. R. (2020). Early prediction of chronic disease using an efficient machine learning algorithm through adaptive probabilistic divergence based feature selection approach. *International Journal of Pervasive Computing and Communications*, *17*(1), 20-36. https://doi.org/10.1108/IJPCC-04-2020-0018

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*. https://doi.org/10.48550/arXiv.2012.12089

Karthick, S. (2017). Semi supervised hierarchy forest clustering and KNN based metric learning technique for machine learning system. *Journal of Advanced Research in Dynamical and Control Systems*, *9*(1), 2679-2690.

Khan, B., Naseem, R., Muhammad, F., Abbas, G., & Kim, S. (2020). An empirical evaluation of machine learning techniques for chronic kidney disease prophecy. *IEEE Access*, *8*, 55012-55022. DOI: 10.1109/ACCESS.2020.2981689

Lakshmanaprabu, S. K., Mohanty, S. N., Krishnamoorthy, S., Uthayakumar, J., & Shankar, K. (2019). Online clinical decision support system using optimal deep neural networks. *Applied Soft Computing*, *81*, 105487. https://doi.org/10.1016/j.asoc.2019.105487

Lambert, J. R., & 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. https://doi.org/10.2174/2666255813999200818131835

Lim, C. C., He, F., Li, J., Tham, Y. C., Tan, C. S., Cheng, C. Y., ... & Sabanayagam, C. (2021). Application of machine learning techniques to understand ethnic differences and risk factors for incident chronic kidney disease in Asians. *BMJ Open Diabetes Research and Care*, *9*(2), e002364. DOI: 10.1136/bmjdrc-2021-002364

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., ... & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, *122*, 56-69.

https://doi.org/10.1016/j.jclinepi.2020.03.002

Onan, A. (2018a). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, *44*(1), 28-47. DOI: 10.1177/0165551516677911

Onan, A. (2018b). Biomedical text categorization based on ensemble pruning and optimized topic modelling. *Computational and Mathematical Methods in Medicine*, 2018. https://doi.org/10.1155/2018/2497471

Onan, A. (2019a). Consensus clustering-based undersampling approach to imbalanced learning. *Scientific Programming*, 2019. doi.org/10.1155/2019/5901087

Onan, A. (2019b). Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access*, *7*, 145614-145633. DOI: 10.1109/ACCESS.2019.2945911

Onan, A. (2019c). Topic-enriched word embeddings for sarcasm identification. In *Computer science on-line conference* (pp. 293-304). Springer, Cham. https://doi.org/10.1007/978-3-030-19807-7_29

Onan, A. (2020). Mining opinions from instructor evaluation reviews: a deep learning approach. *Computer Applications in Engineering Education*, *28*(1), 117-138. DOI: 10.1002/cae.22179

Onan, A. (2021a). Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. *Computer Applications in Engineering Education*, *29*(3), 572-589. DOI: 10.1002/cae.22253

Onan, A. (2021b). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, *33*(23), e5909. DOI: 10.1002/cpe.5909

Onan, A. (2022). Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 2098-2117. doi.org/10.1016/j.jksuci.2022.02.025

Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25-38.DOI: 10.1177/0165551515613226

Onan, A., & Toçoğlu, M. A. (2021). A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification. *IEEE Access*, *9*, 7701-7722. DOI: 10.1109/ACCESS.2021.3049734

Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, *57*, 232-247. https://doi.org/10.1016/j.eswa.2016.03.045

Onan, A., Korukoğlu, S., & Bulut, H. (2017). A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management*, *53*(4), 814-833. https://doi.org/10.1016/j.ipm.2017.02.008

Parmar, M. D., Pang, W., Hao, D., Jiang, J., Liupu, W., Wang, L., & Zhou, Y. (2019b). FREDPC: A feasible residual error-based density peak clustering algorithm with the fragment merging strategy. *IEEE Access*, 7, 89789-89804. https://doi.org/10.1109/ACCESS.2019.2926579

Parmar, M., Wang, D., Zhang, X., Tan, A. H., Miao, C., Jiang, J., & Zhou, Y. (2019a). REDPC: A residual error-based density peak clustering algorithm. *Neurocomputing*, 348, 82-96. https://doi.org/10.1016/j.neucom.2018.06.087

Rady, E. H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*, *15*, 100178. https://doi.org/10.1016/j.imu.2019.100178

Ravindra, B. V., Sriraam, N., & Geetha, M. J. I. J. E. T. (2018). Classification of non-

chronic and chronic kidney disease using SVM neural networks. *International Journal of Engineering & Technology*, *7*(1), 191-194. DOI: 10.14419/ijet.v7i1.3.10669

Scholar, P. G. (2018). Chronic kidney disease prediction using machine learning. *International Journal of Computer Science and Information Security (IJCSIS)*, *16*(4). DOI: 10.35940/ijeat.A2213.109119

Senan, E. M., Al-Adhaileh, M. H., Alsaade, F. W., Aldhyani, T. H., Alqarni, A. A., Alsharif, N., ... & Alzahrani, M. Y. (2021). Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. *Journal of Healthcare Engineering*, *2021*. doi.org/10.1155/2021/1004767

Shetty, A. R., Ahmed, F. B., & Naik, V. M. (2019). CKD prediction using data mining technique as SVM and KNN With pycharm. *International Research Journal of Engineering and Technology (IRJET)*, *6*(5), 4399-4405.

Sobrinho, A., Queiroz, A. C. D. S., Da Silva, L. D., Costa, E. D. B., Pinheiro, M. E., & Perkusich, A. (2020). Computer-aided diagnosis of chronic kidney disease in developing countries: A comparative analysis of machine learning techniques. *IEEE Access*, *8*, 25407-25419. DOI: 10.1109/ACCESS.2020.2971208

Thongprayoon, C., Kaewput, W., Choudhury, A., Hansrivijit, P., Mao, M. A., & Cheungpasitporn, W. (2021). Is It Time for Machine Learning Algorithms to Predict the Risk of Kidney Failure in Patients with Chronic Kidney Disease?. *Journal of Clinical Medicine*, *10*(5), 1121. https://doi.org/10.3390/jcm10051121

Wang, W., Chakraborty, G., & Chakraborty, B. (2020). Predicting the risk of chronic kidney disease (ckd) using machine learning algorithm. *Applied Sciences*, *11*(1), 202. https://doi.org/10.3390/app11010202

Wang, Y., Ding, S., Wang, L., & Ding, L. (2021). An improved density-based adaptive p-spectral clustering algorithm. *International Journal of Machine Learning and Cybernetics*, *12*(6), 1571-1582. https://doi.org/10.1007/s13042-020-01236-x

Wen, G. (2020). Robust self-tuning spectral clustering. *Neurocomputing*, *391*, 243-248. https://doi.org/10.1016/j.neucom.2018.11.105

Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., ... & Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of translational medicine*, *17*(1), 1-13. DOI:10.1186/s12967-019-1860-0

Zelnik-manor, L., & Perona, P. (2004). Self-Tuning Spectral Clustering. *Advances in Neural Information Processing Systems*, *17*, 1-8. https://proceedings.neurips.cc/paper/2004/file/40173ea48d9567f1f393b20c855bb40b-Paper.pdf

Zhang, X., Li, J., & Yu, H. (2011). Local density adaptive similarity measurement for spectral clustering. *Pattern Recognition Letters*, *32*(2), 352-358. https://doi.org/10.1016/j.patrec.2010.09.014

Zhang, Y., Yang, Y., Li, T., & Fujita, H. (2019). A multitask multiview clustering algorithm in heterogeneous situations based on LLE and LE. *Knowledge-Based Systems*, *163*, 776-786. doi.org/10.1016/j.knosys.2018.10.001