

Caviar Waterwheel Plant Optimization-Based Feature Selection For EarlyStage Chronic Kidney Disease Big Data

Mamatha B^{1*} Sujatha P Terdal²

^{1*}Research Scholar Department of CSE PDA College of Engineering Gulbarga, India mamatha.789@gmail.com

²Professor Department of CSE PDA College of Engineering Gulbarga, India sujatha.terdal@gmail.com

Citation: Mamatha B, Sujatha P Terdal (2024), Caviar Waterwheel Plant Optimization-Based Feature Selection For EarlyStage Chronic Kidney Disease Big Data *Educational Administration: Theory And Practice*, 30(5), 1206-1215
Doi: 10.53555/kuev.v30i5.3038

ARTICLE INFO

ABSTRACT

Chronic kidney disease, also termed as chronic renal disease or CKD, is a condition distinguished by a progressive loss in kidneys function over time. Diabetes as well as high blood pressure is the major frequent reason of CKD and sometimes it leads to death if it is left untreated. Several approaches were developed so far to find the CKD but, they failed due to lack of sufficient features and proper selection of features. In order to overcome such gaps, a hybrid optimization-driven model is developed for feature selection of early Stage of CKD utilizing Big data. The input big data is subjected to a data partitioning phase, which is conducted by Bayesian Fuzzy Clustering (BFC). Thereafter, the partitioned data is fed into the Spark architecture which includes two nodes

- a slave node and a master node. After that, pre-processing is carried out by the slave nodes utilizing the z-score normalization approach. Each node pre-processes a portion of the data and sends the output to the master node. The master node merges fused data from all the slave nodes. Finally, feature selection is done based on proposed Conditional Autoregressive Value-at-Risk Water Wheel Plant Algorithm (CWWPA), where fitness is based on entropy. The proposed CWWPA is devised by incorporating the Water Wheel Plant Algorithm (WWPA) and the CAViaR concept. CWWPA achieved maximum correlation coefficient of 0.948 also achieved minimum entropy of 0.399 with cluster size=6.

Key words: Big data, Chronic Kidney Disease (CKD), Bayesian Fuzzy Clustering (BFC), Water Wheel Plant Algorithm (WWPA), CAViaR.

1. Introduction

Big data is a term that was introduced in 1990s for including databases too large to be utilized with general software [12]. Big data is an enormous amount of database, which can be stored, processed, as well as evaluated by the recently accessible tools and database management systems. The employ of big data can be constantly enlarged in various fields, mostly for the principle of diagnosing as well as predicting the disease of the patients [3]. CKD is distinguished by a progressive as well as long-term refuse in function of kidney, which leads to an incapability to efficiently filter waste and maintain balance in fluid as well as electrolyte, which results in of waste products accumulation and retention of fluid [6]. CKD is connected with an enlarged risk of cardiovascular disease (CVD). It is significant to note that CKD is a severe medical condition; it can be cured if it is detected earlier [5]. Presently, risk declaration and prevention of renal failure are mostly concentrated on CKD patients. Nevertheless, the attentiveness rate of early CKD is low; more patients with CKD have no clear influences in early stage, resultant in an excessive rate of missing diagnosis amongst common population [1].

With the aid of prediction analysis, medical experts can estimate biological as well as socio- economical threats and motivate early involvement to prevent chronic diseases [10]. The prediction accuracy of categorization algorithms depend on suitable feature selection algorithms for decreasing the amount of features from the collected data. Feature selection is the process of selecting the most suitable features and eliminating unwanted features [11]. Currently, Machine Learning (ML) and Deep Learning (DL) are the recent techniques to evaluate data as well as predicting consequences [7]. Recently, DL techniques have been utilized to automate the extraction and interpretation of the kidney functions. DL is ML which concentrates on learning to illustrate and build multiple levels. DL needs computational models that consists of numerous layers of storage to obtain data structures of diverse abstraction scales [4]. With the broad application of electronic medical record

systems, particularly the development of regional health information platforms for exchanging data as well as sharing, extensive clinical medical data have been collected, which offers strong data support for medical health research [13] [14]. Extracting and assessing retrospective population data from electronic health record (EHR) big data platforms would mostly enlarge the feasibility of many medical studies in the scope of data availability [1].

The crucial intention is to present CWWPA for selecting feature among CKD patients using big data. The input big data acquired from a dataset is passed to a data partitioning phase carried out by BFC. Thereafter, the partitioned data is fed into the Spark architecture which includes two nodes - a slave node as well as a master node. After that, pre-processing is done by the slave nodes utilizing z-score normalization technique. Each node pre-processes a portion of the data and sends the result to the master node. The master node joins fused data from all the slave nodes. Finally, feature selection is done based on proposed CWWPA, where fitness is based on entropy. The proposed CWWPA will be devised by incorporating the WWPAs with CAViaR. The significant contribution is exposed beneath.

❖ **CWWPA based feature selection for early stage CKD using big data:** CKD is a severe medical condition; it can be cured if it is detected earlier. Here, CWWPA is employed for feature selection to predict early stage CKD among patients. And the devised model is the incorporation of WWPAs and CAViaR.

The organizing of below structure is: Reviewed approaches and its demerits are interpreted in section 2, CWWPA methodology is presented in section 3, CWWPA results are enumerated in section 4 whereas section 5 illustrates conclusion of CWWPA.

2. Motivation

CKD is a major risk concern for growth of CVD, and also raise the risk of cardiovascular mortality. This motivated to propose a new model for predicting CKD by reviewing the difficulties of conventional methods. The techniques and its drawbacks are detailed in this part.

2.1 Literature Survey

Yang, Y., *et al.* [1] developed XG-Boost method for predicting CKD. This model achieved a large number of potential variables to predict the risk of renal failure. But, it failed to develop external validation in multiple diverse disease cohorts and evaluation in clinical trials. Ravikumaran, P., *et al.* [2] introduced Deep Belief Network (DBN) to predict CKD which offered effective outcomes with minimal features along with remarkable connectivity. However, it failed to use a benchmark dataset. Ramani, R., *et al.* [3] devised Hybrid Wrapper and Filter based FS Algorithm (HWFFSA) for CKD prediction. This model increased the speed of processor. Nevertheless, it was complex to deal with missing and noisy values. Elkholy, S.M.M., *et al.* [4] presented Deep Belief Network (DBN) for early prediction of CKD consumed low training time. But, it was difficult for classifying data with missing values.

2.2 Challenges

The experienced demerits by reviewed traditional schemes are described below

- In [1], an XG-Boost method was developed to predict renal failure for chronic disease patients. However, it failed to employ the model by collecting more patients with diverse basic diseases.
- The introduced technique in [3] handled a huge volume of dataset samples effectively. But, it failed to develop a classifier model for dealing with noisy as well as missing values.
- The early prediction of CKD is the most succeeding research in this period. Different techniques as well as approaches are incorporated for reducing the progression of kidney damage. Nevertheless, this is a challenging task owing to complexity of reversing renal damage enlarges steadily with the disease progression.

3. Proposed CWWPA for early stage prediction of CKD

Renal failure is a deadly disease increasing global issues. Earlier risk approaches for renal failure typically rely on the diagnosis of CKD, which not have clear clinical symptoms and thus is generally undiagnosed, which causes high risk among patients. To reduce the risk, an optimization-driven model is developed for the feature selection of early Stage CKD using Big data. The input big data acquired from a dataset [16] is passed to a data partitioning module, done using BFC [15]. Once the data is partitioned, it is fed into the Spark architecture which consists of two nodes - a slave node and a master node. The pre-processing of data is carried out by the slave nodes using the z-score normalization technique. Each node pre-processes a portion of the data and feeds the result to the master node. The master node merges combined data from all the slave nodes. Thereafter, feature selection is done based on proposed CWWPA, where fitness is based on entropy. The proposed CWWPA is devised by incorporating the WWPAs [8] with CAViaR [9]. Figure 1 represents CWWPA for early prediction

of CKD using big data.

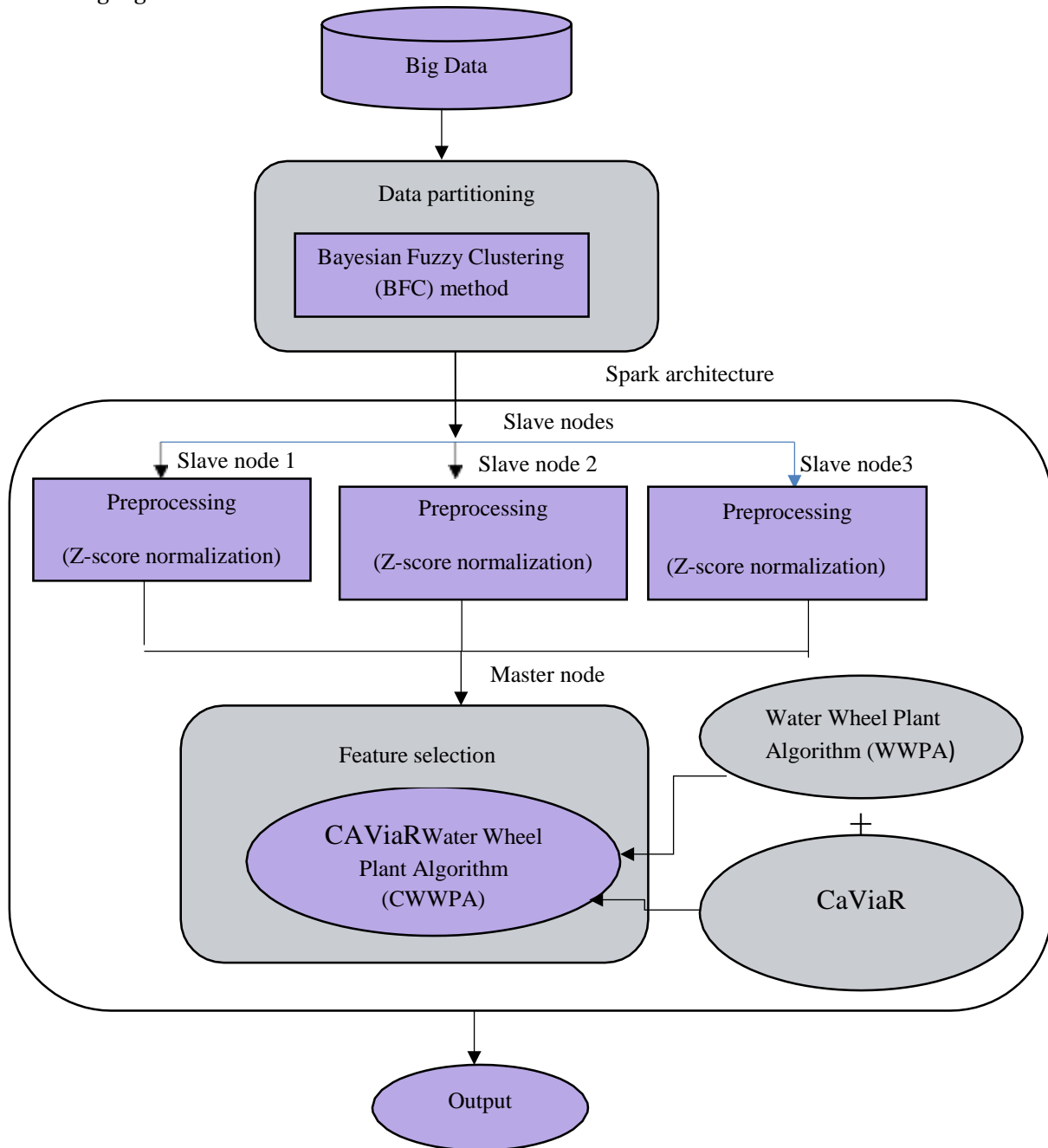


Figure1. Pictorial illustration of CWWPA for early prediction of CKD using big data

3.1 Acquisition of data

The input data is obtained from database that can be defined as below,

$$A = \{A_1, A_2, \dots, A_n\}.(1)$$

Here, A denotes CKD database whereas,

3.2 Data Partitioning utilizing BFC

A_n specifies the number of data

Data Partitioning refers to a process, where the data is divided into several partitions. Here, the input data given to data partitioning phase is A, which is conducted by BFC.

BFC

BFC [15] model is utilized for handling a variable number of clusters and its algorithmic steps are given as below,

Step 1: Initialization

The hyper parameters of this allocation can be placed in the empirical Bayes approach for utilizing the mean of the database, which is defined as follows,

$$\eta_j = \frac{1}{P} \sum_{r=1}^P L_r \quad (2)$$

In addition, a large covariance in the form of the data covariance, which is determined as the below expression,

$$\Sigma_j = \frac{\lambda}{P} \sum_{r=1}^P (L_r - \eta_j)(L_r - \eta_j)^M \quad (3)$$

where, λ is represented as a user set parameter affecting the prior's strength.

Step 2: Evaluating membership vector

For simplicity, a uniform symmetric Dirichlet suggestion allocation was calculated by,

$$a_r^* \sim \text{Dirichlet}(\beta = 1) \quad (4)$$

where, $D(U|L, H)$ denotes conditional distribution of memberships, which is proportional to joint distribution of data, prototypes as well as memberships, $D(L, U, H)$ specified permanent values of the cluster prototypes and a_r denotes membership vector.

A novel membership sample a_r^* is, consequently, accepted to change a_r with probability capable of the ratio is expressed as follows,

$$d_a = \min\left\{1, \frac{\tilde{D}(L_r, a_r^* | H)}{D(L_r, a_r | H)}\right\} \quad (5)$$

Step 3: Evaluating cluster prototype

The new values for cluster prototypes are sampled from a Gaussian distribution through tiny variance in the form of the prior as well as centered on the present Markov-chain state.

$$G_h^* \sim S(G_h, \frac{1}{\rho} \Sigma_h) \quad (6)$$

Here, ρ represents user set parameter, which controls tightness around the present state as well as relates to the sample acceptance rate.

The sample can be accepted in accordance with the ratio is calculated by using the following expression.

$$d_a = \min\left\{1, \frac{\tilde{D}(L, G_h^* | H)}{D(L, G_h | H)}\right\} \quad (7)$$

Here, the Gaussian proposal is symmetric. The result obtained by BFC is the cluster data which is denoted as, $K = \{K_1, K_2, K_3\}$.

3.3 Spark Architecture

The spark architecture comprises of a single master and multiple slaves. Here, the inputs passed to spark architecture are the clustered data $K = \{K_1, K_2, K_3\}$ to the slave nodes SN_1, SN_2 , and SN_3 . The process carried out by

Slave node

SN_1 is described as follows,

The data nodes are referred as slaves which distribute the task belongs to the submitted job. The slave nodes

transmit data after getting a request from master node. The slave nodes did not exchange a data with each other.

a) Pre-processing using Z-score normalization

Pre-processing is the model of altering unprocessed data into a clean data base. The database is preprocessed for checking missing values, noisy data, and other inconsistencies and makes it suitable for further process. Here, the pre-processing is carried out by Z-score Normalization. It isa process of normalizing each value in a database in such a manner the mean of all values is 0 as well as standard deviation is1.Let m_j ($j=1,2,\dots, R$) represents the j^{th} component of every feature vector $I \in \mathbb{N}^R$ and calculate the mean as well as standard deviation of these R components:

$$\gamma_I = \frac{1}{R} \sum_{g=1}^R m_g, \quad \mu_I = \sqrt{\frac{1}{R} \sum_{g=1}^R (m_g - \gamma_I)^2} \tag{8}$$

After applying Z-score normalization the expression becomes.

$$I^{zn} = ZN(I) = \frac{I - \gamma_I \mathbf{1}}{\mu_I} \in \mathbb{N}^R \tag{9}$$

Here, $\mathbf{1}=[1, 1, \dots,1]^T$ represents a R – dimensional vector and its components being ones.

From the above expression the z-score feature normalization, obtained that,

$$\| I^{(zn)} \|^2 = \sum_{g=1}^R \left(\frac{m_g - \gamma_I}{\mu_I} \right)^2 \tag{10}$$

$$\| I^{(zn)} \| = \sqrt{\frac{\sum_{g=1}^R (m_g - \gamma_I)^2}{R}} = \sqrt{R} \tag{11}$$

$$\langle I^{(zn)}, \mathbf{1} \rangle = \sum_{g=1}^R \left(\frac{m_g - \gamma_I}{\mu_I} \right) \cdot 1 \tag{12}$$

$$\frac{\sum_{g=1}^R m_g - R\gamma_I}{\mu_I} \tag{13}$$

Here, $\langle \dots, \dots \rangle$ specifies dot product of two vectors. The computation of z-score normalization first projects the unique feature vectors along $\mathbf{1}$ vector to a hyper- plane that consists of the origin aswell as is perpendicular to $\mathbf{1}$. These vectors are then scaled to the similar length of R , that is the finishing normalized vectors lie on a hyper- sphere with radius \sqrt{R} . The result obtained from this phase is the normalized data. Here, the result from Pre-processing phase of $SN_1 B_1$ having $n \times m$ isdimension. Thus, the same process is executed for the remaining slave nodes.

Master node

It is a way of asymmetric communication where the master controls one or more other devices orprocesses is termed as slaves and serves as their communication hub. Merging data for the entire slave node and fed it to the feature selection phase for obtaining the detected results. The input passed to master node is B, such that $B \in \{B_1, B_2, B_3\}$.

Solution encoding

Solution encoding refers to select the best features. Figure 2 shows the solution encoding. Thus, the optimal solution is defined as below,

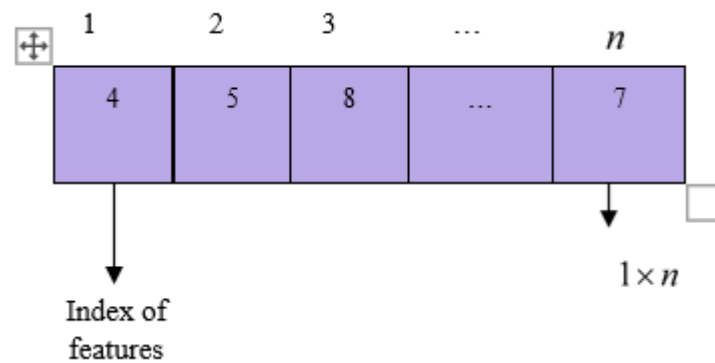


Figure 2. Solution encoding Here, n represents the total features to be selected.

Fitness

The fitness is calculated based on entropy. Entropy computes the information comprises in the segmented parts as well as the total amount of disorders present in the system. It can be computed by using the following expression,

$$N(\underline{Z}) = \sum_{\underline{Z}} U(\underline{Z}) \log \frac{1}{U(\underline{Z})} \quad (14)$$

Here, Z implies the candidate features.

Feature selection using CWWPA

WWPA is a stochastic optimization approach motivated by natural systems. The basic concept of WWPA is based upon modeling the waterwheel plant's natural behavior while on hunting expedition. A CAViaR model specifies the progression of the quantile over time utilizing a special kind of autoregressive process. CWWPA is developed by combining WWPA and CAViaR, which is capable for addressing optimization problems and also it is utilized for selecting best features. The algorithm of CWWPA is expressed as follows,

Step 1: Initialization

Let us initialize the position of waterwheels in a search space at a random manner, which can be calculated by,

$$Q = \{Q_1, Q_2, \dots, Q_v, \dots, Q_w\} \quad (15)$$

Here, Q specifies the population, whereas Q_v denotes standard solution and Q_w signifies the total number of solution.

$$Q_{v,k} = Cx_k + y_{v,k} \cdot (Ox_k - Cx_k), v=1,2,\dots,T, k=1,2,\dots,p \quad (16)$$

Here, T represents number of waterwheels as well as p denotes number of variables, $y_{v,k}$ specifies random number in the interval [0.1], Cx_k denotes lower bound of k^{th} problem variable, Whereas Ox_k signifies upper bound of k^{th} problem variable, $Q_{v,k}$ signifies v^{th} solution in k^{th} dimension.

Step 2: Computation of Objective function

The objective function is computed based on entropy using Eq. (14).

Step 3: Position Identification and Hunting of Insects

To demonstrate the new position of the waterwheel, the expression below is utilized in conjunction with the simulation of the waterwheel's technique to the insect.

$$Q(s) = \psi_0 + \sum_{i=1}^e \psi_i Q(s-i) + \sum_{l=1}^v B(Q(s-l)), \quad \text{where } e=v=2 \tag{17}$$

The standard CAViaR is expressed as follows,

$$Q(s) = \psi_0 + \sum_{i=1}^e \psi_i Q(s-i) + \sum_{l=1}^v B(Q(s-l)), \quad \text{where } e=v=2 \tag{18}$$

$$Q(s) = \psi_0 + \psi_1 Q(s-1) + \psi_1 Q(s-2) + \psi_1 B(Q(s-1)) + \psi_2 B(Q(s-2)) \tag{19}$$

Substitute Eq. (19) in Eq. (17), then the updated expression of CWWPA becomes,

$$Q(s+1) = \psi_0 + \psi_1 Q(s-1) + \psi_1 Q(s-2) + \psi_1 B(Q(s-1)) + \psi_2 B(Q(s-2)) + \mathfrak{R} E \tag{20}$$

Here, \mathfrak{R} represents exponential variable in range $[0,1]$, t_1 and t_2 specifies random variable in interval $(0,2)$ and $(0,1)$ as well as E specifies the candidate features.

Step 4: Carrying the Insect in the Suitable Tube

If the target function value is superior at this new site, the waterwheel is moved in the place of prior position, as shown in the below equation,

$$Q(s+1) = Q(s) + \mathfrak{R}E \tag{21}$$

Step 5: Termination

The above steps are performed continuously until achieving an optimal solution.

4. Result and Discussions

The results of CWWPA to prove the efficiency is discussed in this segment.

Experimental setup

CWWPA devised for predicting CKD utilizing big data is trained in PYTHON tool.

Dataset Description

Chronic kidney disease dataset [16] is a multivariate database. It consists of 400 instances and the feature type is real.

Assessment metrics

The metrics employed for assessment of CWWPA are correlation coefficient as well as entropy.

• Correlation coefficient

The correlation coefficient refers to a statistical measure of strength of a linear difference among two variables.

$$\omega = \frac{\partial(\sum TN) - (\sum T)(\sum N)}{\sqrt{[\partial^*(\sum T^2 - (\sum T)^2)] * [(\sum N^2 - (\sum N)^2)]}} \tag{22}$$

Here, ω denotes correlation coefficient, whereas ∂ represents number of observation and the variables are denoted as T and N.

• Entropy

Entropy is an evaluation of randomness or disorder of a system, which is calculated using Eq. (14).

Comparative approaches

XG-Boost [1][17], DBN [2], HWFFS [3] and DBN [4] are the approaches assumed for assessment of CWWPA for revealing its effectiveness.

4.5 Comparative assessment

The comparative assessment of CWWPA for CKD prediction employing big data is accomplished for three setups with cluster size= 3, 4 and 6 by varying number of features.

4.5.1 Evaluation with cluster size 3

Figure 3 exhibits calculation regarding CWWPA by altering number of features for cluster size=3. Figure 3 a) represents assessment of CWWPA in terms of correlation coefficient. CWWPA attained correlation coefficient of 0.817, whereas XG-Boost, DBN, HWFFS and DBN attained 0.709, 0.729, 0.759 and 0.777 for number of features=20. Evaluation of CWWPA in regards of entropy is interpreted in figure 3 b). XG-Boost, DBN, HWFFS and DBN gained entropy of 0.788, 0.760, 0.710 and 0.678 while, CWWPA obtained 0.519 while assuming the number of features=20.

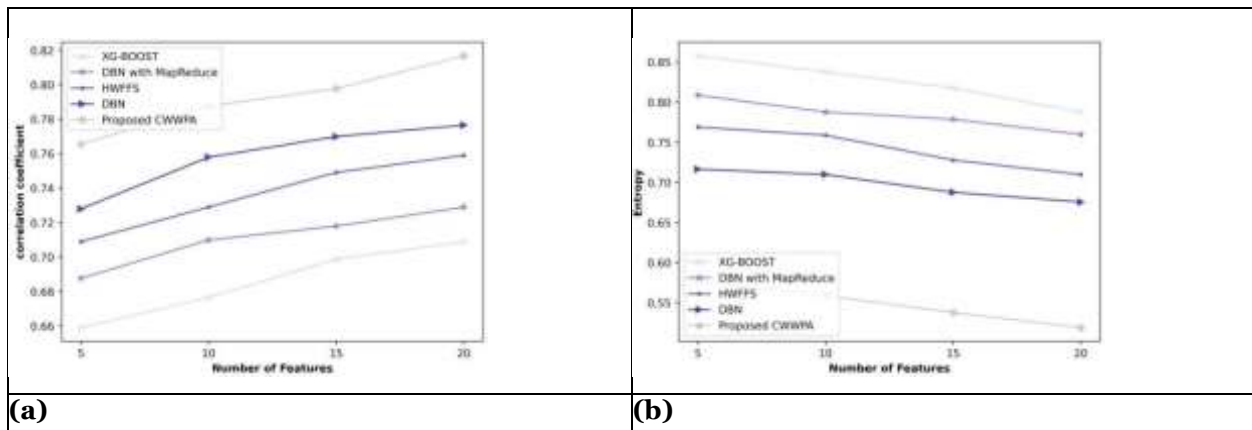


Figure 3. Analysis of CWWPA using cluster size-3, a) Correlation coefficient, b) Entropy

4.5.2 Assessment with cluster size 4

Figure 4 illustrates estimation with respect to CWWPA by varying number of features for clustersize=4. Figure 4 a) deliberates assessment of CWWPA by means of Correlation coefficient. Correlation coefficient attained by CWWPA is 0.908 while, value obtained by XG-Boost is 0.730, DBN is 0.749, HWFFS is 0.770 and DBN is 0.799 for number of features as 20. Entropy is elucidated in figure 4 b). The gained entropy of CWWPA, while assuming the number of features as 20 is 0.405.

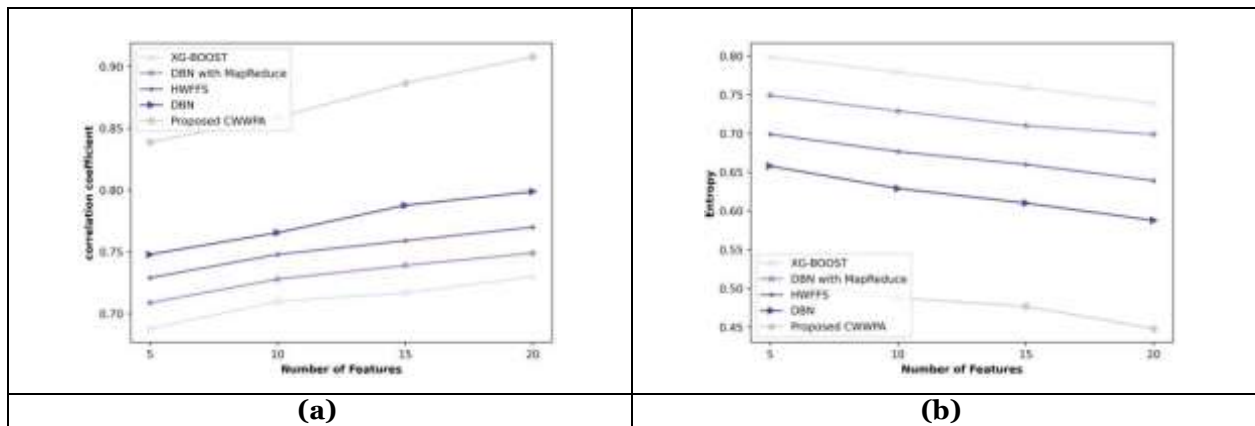


Figure 4. Analysis of CWWPA using cluster size-4, a) Correlation coefficient, b) Entropy

4.5.3 Estimation with cluster size 6

Figure 5 represents analysis in regards of CWWPA by altering number of features for cluster size=6. Figure 5 a) represents evaluation of CWWPA by means of correlation coefficient. CWWPA obtained correlation coefficient of 0.948, whereas XG-Boost, DBN, HWFFS and DBN attained 0.769, 0.788, 0.838 and 0.877 for 20th number of features. Assessment of CWWPA with respect to entropy is 0.488, whereas XG-Boost, DBN, HWFFS and DBN attained entropy as 0.629, 0.559, 0.488 and 0.428 for number of features=20 as mentioned in figure 5 b).

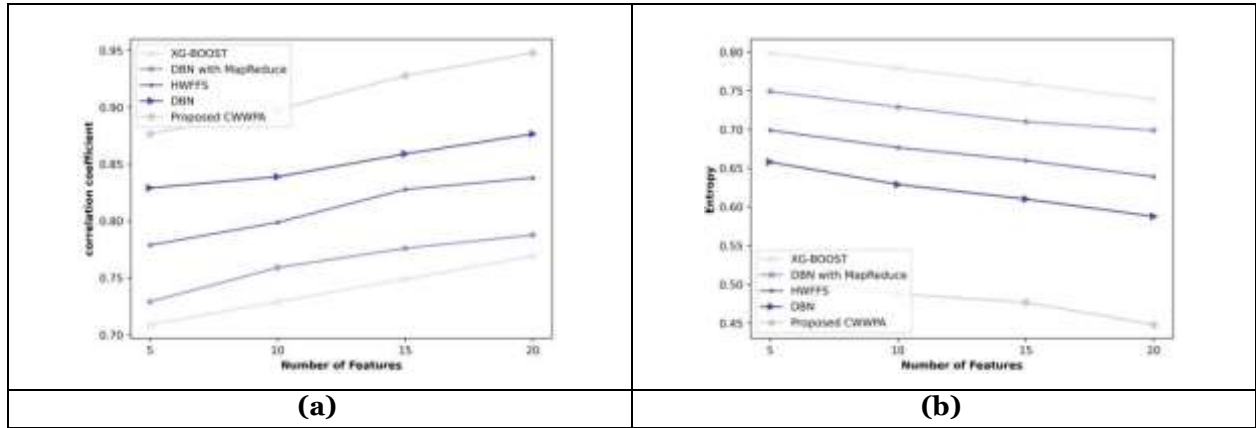


Figure 5. Analysis of CWWPA using cluster size-6, a) Correlation coefficient, b) Entropy

4.6 Comparative discussion

Table 1 signifies comparative discussion and values obtained by developed and traditional techniques during evaluation. CWWPA obtained maximum correlation coefficient of 0.948 also achieved minimum entropy of 0.399 with cluster size=6.

Table 1. Comparative discussion of CWWPA

Analysis based upon	Metrics/Methods	XG-Boost	DBN	HWFFS	DBN	Proposed CWWPA
Setup with cluster size =3	Correlation coefficient	0.709	0.729	0.759	0.777	0.817
	entropy	0.788	0.760	0.710	0.676	0.519
Setup with cluster size =4	Correlation coefficient	0.730	0.749	0.770	0.799	0.908
	entropy	0.668	0.633	0.597	0.588	0.537
Setup with cluster size =6	Correlation coefficient	0.769	0.788	0.838	0.877	0.948
	entropy	0.629	0.559	0.488	0.428	0.399

5. Conclusion

CKD has a significant impact on the mortality as well as morbidity of patients. CKD often coexists with CVD as well as diabetes and is acknowledged as a threat factor for all-cause mortality and cardiovascular disease. Initially, input big data acquired from a database is passed to a data partitioning phase, which is conducted by BFC. Thereafter, the partitioned data is subjected to the Spark architecture which includes two nodes - a slave node and master node. After that, pre-processing is carried out by the slave nodes utilizing the z-score normalization approach. Each node pre-processes a portion of the data and sends the output to the master node. The master node combines fused data from all the slave nodes. Finally, feature selection is done based on proposed CWWPA, where fitness is based on entropy. The proposed CWWPA is devised by incorporating the WPPA with CaViaR. CWWPA obtained maximum correlation coefficient of 0.948 and minimum entropy of 0.399 with cluster size=6. The future scope would be the inclusion of classification of CKD step as well as severity level categorization of such disease.

References

[1] Yang, Y., Li, Y., Chen, R., Zheng, J., Cai, Y. and Fortino, G., "Risk prediction of renal failure for chronic disease population based on electronic health record big data", Big Data Research, vol.25, pp.100234, 2021.

[2] Ravikumaran, P., Devi, K.V. and Valarmathi, K., "Prediction of chronic and non-chronic kidney disease

- using modified DBN with map and reduce framework", pp. 185-202, 2022.
- [3] Ramani, R., Devi, K.V. and Soundar, K.R., " MapReduce-based big data framework using modified artificial neural network classifier for diabetic chronic disease prediction", *Soft Comput.*,vol. 24, no. 21, pp. 16335-16345,2020.
- [4] Elkholy, S.M.M., Rezk, A. and Saleh, A.A.E.F., "Early prediction of chronic kidney disease using deep belief network", *IEEE Access*, vol. 9, pp. 135542-135549,2021.
- [5] Ismail, W.N., "Snake-Efficient Feature Selection-Based Framework for Precise Early Detection of Chronic Kidney Disease", *Diagnostics*, vol.13, no. 15, pp.2501, 2023.
- [6] Arif, M.S., Mukheimer, A. and Asif, D., " Enhancing the Early Detection of Chronic Kidney Disease: A Robust Machine Learning Model", *Big Data and Cognitive Computing*, vol.7, no. 3, pp. 144,2023.
- [7] Samrat Kumar Dey, Khandaker Mohammad Mohi Uddin, Hafiz Md. Hasan Babu, Md. Mahbubur Rahman, Arpita Howlader, K.M. Aslam Uddin,"Chi2-MI: A hybrid feature selection based machine learning approach in the diagnosis of chronic kidney disease", *Intelligent Systems with Applications*,Vol.16,2022.
- [8] Abdelhamid, A.A., Towfek, S.K., Khodadadi, N., Alhussan, A.A., Khafaga, D.S., Eid, M.M. and Ibrahim, A., "Waterwheel Plant Algorithm",*A Novel Metaheuristic Optimization Method. Processes*, vol.11, no.5, pp.1502, 2023.
- [9] Engle, R.F. and Manganelli, S., "CAViaR: Conditional autoregressive value at risk by regression quantiles", *Journal of Business& Economic Statistics*, vol. 22,no. 4, pp. 367-381, 2004.
- [10] Ahmed, K., Singh, S., Pathak, C., And Singh, S.," Big Data Analytics For Chronic Disease Prediction" Vol.4, 2022.
- [11] Sara, S.B.V. and Kalaiselvi, K., "Ensemble swarm behavior based feature selection and support vector machine classifier for chronic kidney disease prediction", *International Journal of Engineering & Technology*, vol. 7, no. 2, pp. 190, 2018.
- [12] Mallappallil, M., Sabu, J., Gruessner, A. and Salifu, M., "A review of big data and medical research",*SAGE open medicine*, vol. 8, pp. 20503121-20934839,2020.
- [13] Safran, C., Bloomrosen, M., Hammond, W.E., Labkoff, S., Markel-Fox, S., Tang, P.C. and Detmer, D.E., "Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper", *Journal of the American Medical Informatics Association*, vol. 14, no. 1, pp. 1-9, 2007.
- [14] Roski, J., Bo-Linn, G.W. and Andrews, T.A.,"Creating value in health care through big data:opportunities and policy implications", *Health Affairs*, vol. 33, no. 7, pp. 1115-1122, 2014.
- [15] Glenn, T.C., Zare, A. and Gader, P.D., "Bayesian fuzzy clustering", *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 5, pp.1545-1561, 2014.
- [16] The Chronic_Kidney_Diseasedatabase will be taken from"https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease#", accessed on October 2023.
- [17] B. Mamatha and S. P. Terdal, "Predicting Chronic Kidney Disease using Machine Learning in the Early Stages," 2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS), Kalaburagi, India, 2023, pp. 1-8, doi: 10.1109/ICIICS59993.2023.10420988.
- [18] Mamatha, B and Sujatha P Terdal. "A Review On Early Detection Of Chronic Kidney Disease." *Journal of Scientific Research and Technology* (2024): n. page 35-43.
- [19] Ranjana B Nadagoudar,et," A Comparative Analysis of Machine Learning Techniques for DGA based botnet detection" *International Journal for Research in Engineering Application & Management (IJREAM) ISSN : 2454-9150 Vol-07, Issue-12, MAR 2022*