OPTIMIZED UNCERTAINTY HANDLING FEATURE SELECTION USING MOTH FLAME BASED DBSCAN CLUSTERING FOR DIABETIC RELATED CHRONIC KIDNEY DISEASE

P. USHA[1], N. KAVITHA[2]

[1]Research scholar, Dept. of Computer Science, Nehru arts and science college
/Assistant Professor Dr. N. G. P. Arts and Science College, Coimbatore, India
[2]Associate Professor & Head, Dept. of Computer Science,
Nehru Arts and Science college, Coimbatore, India
Email: blessed.usha@gmail.com

ABSTRACT

In this work, to overcome the problem of voluminous dataset handling, improving the relevancy of features to detect the chronic kidney disease related diabetic patients, an uncertainty handling feature selection is constructed. The dataset is initially clustered by DBSCAN, the efficiency of clustering is achieved by two parameters they are epsilon and min-points. These are assigned in a random manner in standard DBSCAN, if the value is not assigned with any knowledge, then the clustering will not produce appropriate result. Hence, this presented work adapted the Moth Flame Optimization to search for optimized parameter values using its flight path behavior It discovers the best fittest value based on the flame position and those are fine tuned. The feature's linear correlation among the class variable is identified using Pearson's linear correlation based on it's the features which doesn't belong to any cluster are eliminated and based on the correlation score they are ranked. Thus, the significant feature subset is used for chronic kidney disease detection. The uncertainty about the outliers is well handled using the moth flame metaheuristic model by finding the best fittest and worst fittest features. The simulation results also proved that the accuracy obtained by the proposed Moth Flame based DBSCAN along with Pearson Correlation (MFDBSCAN-PC) produced highest accuracy compared to other standard feature selection models in chronic kidney disease detection with reduced error rate.

Keywords: *Chronic kidney disease, Density based Spatial Clustering along with Noise, Feature selection, Moth Flame, Metaheuristic model, Pearson Correlation*

1. INTRODUCTION

Advancement in the field of information technology disease diagnosis have greatly improved the healthcare domain. Insistent, low-grade information is now considered as an essential feature for chronic kidney disease detection (CKD). From the reports it stated that nearly 10 to 15% of the population is affected by critical CKD health problem and its rifeness is unceasingly growing [1]. The victims at its CKD early stage, does not have any major symptoms and it is work difficult to discover without few of the test like blood and urine test. When the presence of CKD is dragonized at its primary stage, defensive actions and improved treatment can be given to control the transplantation or chances of dialysis [2].

Data mining approaches plays a vital role in prediction of CKD at its earlier stages, using its various algorithms like classification, clustering, regression, etc. But to use them in a better manner it is very important to handle the voluminous data which should not be influenced by the irrelevant and redundant features [3]. Feature selection is an essential part of the mining approach, which involves in eliminating the redundant and irrelevant features by discovering features which will improve the accuracy of the algorithm and reduce the complexity of computation.

Hence, in this work CKD dataset is collected from the UCI Machine learning repository, with huge volume of attributes. To produce optimized classification of chronic kidney disease it is necessary to detect the importance of each attribute. Because not all the features will contribute to the detection process, so a novel optimized model is developed in this proposed research work for finding potential features in the chronic kidney disease dataset using moth flame optimization-based density clustering and ranking the features by computing the linear correlation with the target variables. Therefore, it improves the relevancy of feature selection and it avoids the local optima in searching process with the proposed model in an optimized manner.

2. RELATED WORK

Mohamed et al [4] in their work designed an intelligent model which involves in eliminating the redundant and irrelevant features using density-based feature selection, before developing Ant colony-based classification model for chronic kidney disease detection.

Dilli et al [5] anticipated a novel weighted average-based ensemble learning model to impute the missing values in CKD dataset. The presence of missing value will affect the precision rate of the detection model. They performed the missing value imputation with various stages of CKD.

Smita et al [6] devised a novel clustering model which solves the dimensionality issue by combining the correlation measure to generate a good feature subset. The irrelevant features are avoided by applying k-means clustering. The unique features are discovered by correlation measure from each cluster.

Pavya et al [7] in their work F-score and recursive feature elimination model which belongs to the filter-based and wrapper-based feature selection respectively. Thyroid disease is detected using three different steps such as feature extraction, selection and classification. They also used principal component analysis for dimensionality reduction.

Nag Anjaneyulu et al [8] introduced a novel probabilistic based feature selection to enhance the disease classification. The training datasets are collected from real-time patients and new type of disease is detection using the patient disease forecasting.

Rajathi et al [9] presented the K-Nearest Neighbor algorithm and Ant Colony Optimization, an integrated approach is proposed to predict Rheumatic Heart Disease. This work integrated approach has two phases. KNN classification is used and ACO is used to prepare the population and search the optimized algorithm.

Sultana et.al [10] handled the issue of prediction of heart disease using four different classification model to extract hidden information which plays an important role in the decision making. This research utilized the various data mining techniques to envisage the heart disease with more accuracy.

Paul et.al. [11] designed a Dynamic Multi Swarm and Particle Swarm Optimization is used to predict the multifaceted attributes and support the diagnosis of heart disease. It involves in noise removal, construction of the rules using fuzzy rule.

3. MATERIALS AND METHODS

This phase focuses on developing an optimized feature selection model to detect the most potential attributes involved in prediction of diabetic related chronic kidney disease. The Chronic Kidney Disease dataset used in this work is collected from UCI Machine Learning Repository. This dataset is having 24 attributes such as Age, BP, Specific Gravity, Sugar, Blood Glucose, Albumin, White blood cell count, red blood cells, Bacteria, Sodium, Potassium, Serum Creatinine, Blood Urea, Hemoglobin, Red Blood Cell Count, Coronary Artery Disease, Packed Cell, Pus Cell clumps, Pus Cell, Diabetes Mellitus, Anemia, Diabetes Mellitus, Appetite, Anemia and class.

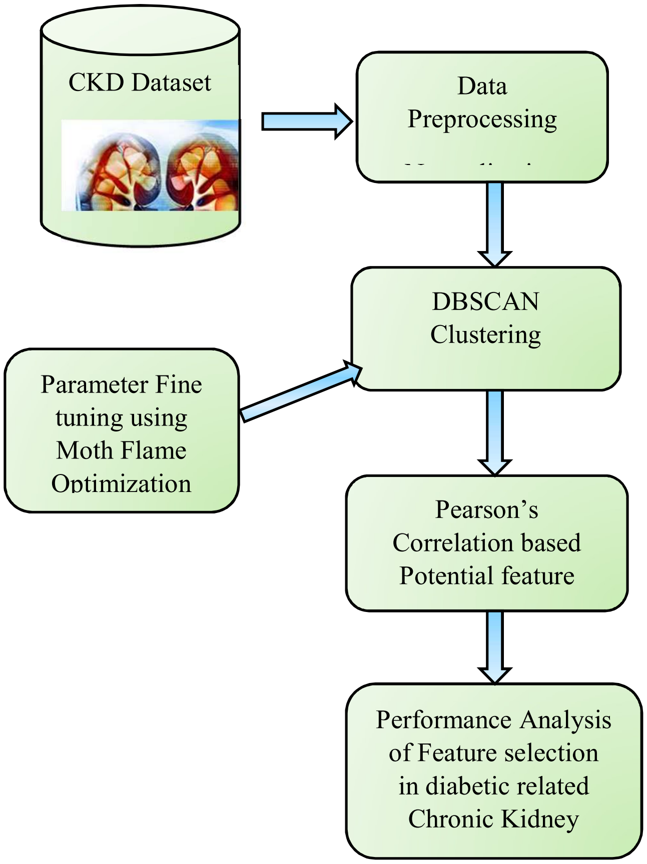
CKD Dataset
Data Preprocessing
DBSCAN Clustering
Parameter Fine tuning using Moth Flame Optimization
Pearson's Correlation based Potential feature
Performance Analysis of Feature selection in diabetic related Chronic Kidney

Figure 1: Significant Feature Selection using Moth Flame based DBSCAN Clustering for Diabetic related chronic kidney disease

This work constructs a Multiclass label feature selection, by clustering the dataset using fuzzy Density based spatial clustering algorithm and then the important attributes which highly influence the formation of clustering is identified using Pearson's correlation-based feature selection and those are used for predicting diabetic related Chronic Kidney disease Prediction.

3.1 Data Preprocessing

The Chronic Kidney Disease Dataset comprised of several attributes with varying range of values. While using the data as such then the higher value attributes are more influenced by the detection process which affects the performance of the model. To overcome this issue and to offer equal importance to all the attributes, its range of value is converted to a common range of interval which falls between 0 to 1. This process is done by min-max normalization and it is mathematically denoted as

$$\text{Norm(D)}=\frac{I_j-\min(I_{j=1\ldots n})}{\max(I_{j\ldots n})-\min(I_{j\ldots n})} \qquad (1)$$

Where D refers to the dataset and I_j refers to the I^{th} instance and its j^{th} attribute, min and max are the minimum and maximum value within the attribute.

3.2. DBSCAN Clustering

While using partition clustering methods and hierarchical clustering discovers spherical shaped clusters, which are suitable only for the compact separated clusters. But they are unable to treat the noise and outlier presented in the dataset. In real time, the clusters might be of random shape and often noise will be presented in the data. To handle such real time restrictions Density Based Spatial Clustering of Applications with Noise (DBSCAN) is appropriate and perfect clustering model. Standard DBSCAN groups the objects depending on their local densities the nearest instances which are closer to each other. Some of the instances which are far in distance are considered as outliers and does not belong to any cluster. It uses two variables epsilon (Eps) is the circle radius to be constructed around each instance to verify their density and midpoints (mP) refers to minimum instances required inside that circle to be classified as core point or centroid, border or outlier

3.3. Algorithm for DBSCAN

- Initialize the value for EPS and mP
- Discover all the neighbor instances within eps and determine the core points with more than mP number of neighbors
- For every core point if it does not belong to a cluster, then create a new cluster
- Identify iteratively all its density linked instances and allocate them to the identical cluster as the core point
- Repeat the process for the remaining instances which has to be clustered. The instances which do not fit to the any cluster is termed as noise or outlier.

In conventional DBSCAN, it is very sensitive to its parameter, if the values are assigned improperly then all the instances become central point or none of the instance will be core point. To overcome this issue, estimation of the parameter's is accomplished in an optimized manner by applying the metaheuristic-based Mouth Flame Optimization algorithm to fine tune the values assigned to the DBSCAN.

3.4. Mouth Flame based DBSCAN Optimization

The Mouth Flame optimization is developed based on the biological behavior of the moths flighting flames in nature. They use a distinct kind of route-finding method for lateral alignment during night flight. The moth flies based on the moon light with a fixed angel maintenance. It updates its position by spiraling around the flame. It is a very simple and effective algorithm as it has a smaller number of parameters involved in finding the optimized solution.

It is assumed that moth is a candidate solution and position is the variable to be solved in its space. By altering their position vectors, moths can fly in more than one dimensions. Moth belongs to a swam intelligence algorithm which is mathematically represented in a matrix as

$$MH = \begin{pmatrix} mh_{1,1} & \cdots & mh_{1.d} \\ \vdots & \ddots & \vdots \\ mh_{n,1} & \cdots & mh_{n,d} \end{pmatrix}$$

Where n denotes number of moths and control variables are represented using d. The list of fitness value vectors related to each moth is embodied as shown

$$FM = \begin{pmatrix} FM_1 \\ FM_2 \\ \vdots \\ FM_n \end{pmatrix}$$

Where FM denotes the fitness value of each corresponding moths in the population. Here, each moth is essential to update their positions only with the unique flame related to it, in order to avoid the local optima value, which will improve the global searching ability of the algorithm. Hence, the moth and flame position in the search space are considered as variable matrices of the identical dimension.

$$FL = \begin{pmatrix} fl_{1,1} & \cdots & fl_{1.d} \\ \vdots & \ddots & \vdots \\ fl_{n,1} & \cdots & fl_{n,d} \end{pmatrix} \qquad (2)$$

Where FL denotes the flame matrix of all the moth's corresponding to its position. The flames fitness values are also matrix which is represented as

$$FF = \begin{pmatrix} FF_1 \\ FF_2 \\ \vdots \\ FF_n \end{pmatrix} \qquad (3)$$

The update strategy during each iteration is different for both moth and flame, this is because the moths truly search individuals that move within the search space, whereas the flame is the best position that iteratively optimized moths can achieve so far. Every individual moth surrounds a related fame, and when a better solution is found, its location is updated to the flame location in next generation. Using this mechanism, MFO able to discover global optimum solution. The mathematical representation of moth to a flame flight behavior and the mechanism to update each moth position related to a flame is also implied as shown below

$$MH_i = L(MH_i - FL_j) \qquad (4)$$

where MH_i signifies ith moth, FL_j refers of the jth flame and the helical function is denoted as L.

The helical function gratifies the subsequent criteria as listed:

- ✓ Initial space of the moth position is chosen as the helical function's initial point
- ✓ Depending on the contemporary flame is treated as the space position of the spiral's end point.

Based on the conditions mentioned the flight path of moth using helical function is represented as

$L(MH_i, FL_j) = d_i * e^{cr} * \cos(2\pi r) + FL_j (5)$

$r = (g - 1) * rnd + 1$ (6)

$g = -1 + itr * \left(-\frac{1}{T_{mx}}\right)$ (7)

Where 'd$_i$' refers to the linear distance among ith moth and jth flame, 'c' refers to helix shape logarithmic constant and 'r' refers to the path coefficient lies between [-1, 1]. The coefficient r position of the moth and the flame in the next iteration, if r = -1 then it denotes that the moth is closest to the flame, else if r = 1 denotes moth is far away from the flame. Its scale is computed as shown in the equation (). The distance d$_i$ is computed as shown

$$d_{i=}FL_j - MH_i \quad (8)$$

It is observed from the equation that the next move of the moth is surrounded by the flame rather than unbiased in the space among them. Thus, it achieves ability of global and local search in optimized manner.

Feature Selection using Mouth Flame based DBSCAN Optimization

Input: Chronic Kidney Disease Dataset (CKD = K$_1$, K$_2$, K$_3$, Kn)

```
Output: PF // Potential feature subset

Algorithm

Stage 1:

//Cluster the instances of CKD dataset using Mouth Flame based DBSCAN Clustering

Assign the parameters for eps and mP using Moth Flame Optimization

    ✓ Initialize the size of moth population as m, $R_{max}$ is the maximum iteration.
    ✓ Set the moth position
    ✓ Compute the fitness value of each moth

          $FM = \begin{pmatrix} FM_1 \\ FM_2 \\ \vdots \\ FM_n \end{pmatrix}$   (9)

    ✓ Sort all Moth based on the fitness value
    ✓ NF = population of moth
    ✓ While ($r < R_{max}$) do

        o Flame Number is updated using equation
          $FL_{nu} = round(NF - r * \frac{NF-1}{R_{max}})$
        o Distance between ith moth with corresponding Flame j is computed
              $d_{i=}FL_j - MH_i$
        o Update the parameters g and r

          $r = (g-1) * rnd + 1$
          $g = -1 + itr * \left(-\frac{1}{T_{mx}}\right)$
          $MH_i = L(MH_i - FL_j)$   (11)
        o The position of moth is updated related to flame ($FL_j$)using the helical spiral function
          $L(MH_i, FL_j) = d_i * e^{cr} * \cos(2\pi r) + FL_j$
        o Update and sort all the moth search agents
        o Update the flames
        o r = r + 1
    ✓ End While

    Stage 2:

    • Determine all the neighbor instances within eps and assign features are core point and determine with more than mP number of neighbors
    • For every core point (feature) if it does not belong to a cluster, then create a new cluster
    • By identifying the density assign the features to the closely related cluster
    • Repeat the process for the remaining features which has to be clustered.
    • The features which do not fit to the any cluster is considered as irrelevant and remove from the dataset
    • Compute the correlation among features using the formula
        o For i = 1 to m

    PC (ft, cl) = $\frac{Exp[(ft-\mu_{ft})((cl- \quad _{cl})]}{\sigma ft \sigma cl}$ (12)

          Where ft is the feature of the dataset, cl is the class variable, $\mu_{ft}$ – mean of feature value, $\mu_{cl}$ – $means\ of\ class\ variable$, standard deviation of feature and the class is denoted by $\sigma ft\ and\ \sigma cl$ respectively.

              ▪ If Pc = 1 remove one of feature
              ▪ PF = $ft_i$
        o End for
    End
```

From the algorithm, the process of effective feature selection using Moth Flame DBSCAN optimization clustering with Pearson's correlation [17] is explained. After the initialization of the moth, during next generation based on the flame the moth will update its position, the first moth usually updates its position with the flame relative to the best fitness value. The last moth will move based on the worst fitness value flame. Once the clustered are generated, the features which doesn't belong to any cluster are eliminated. The next step is to remove redundant features from every cluster. To perform this Pearson correlation is used to measure the correlation among the feature ft and the class variable cl. If PC (ft,cl) = 0 then they are not correlated and if greater than or equal to 1 the feature and the class are linearly dependent. The feature list is ordered by decreasing value of PC (ft, cl) which serves as the ranking of features.

4. RESULTS AND DISCUSSIONS

This section explains in detail about the performance of the proposed Moth Flame based DBSCAN Optimization with Pearson and correlation-based feature selection (MFDBSCAN-PC) to enhance the process of chronic kidney disease detection. The MFDBSCAN-PC is implemented using python code. The Chronic Kidney disease dataset is collected from UCI machine learning repository with 25 attributes with 400 instances [16]. The performance of the proposed MFDBSCAN-PC is compared with the existing three feature selection models Best First Search (BFS), Principal Component Analysis (PCA), Pearson Correlation Feature Selection (PCS) and whole attributes is also considered. The evaluation metrics used for analyzing the performance are Accuracy, Precision, Recall, F-Measure and Error Rate

Table 1: Number of Feature Selected using three different feature selection algorithms

Feature Selection Methods	No. of. Features
Best First Search	18
Principal component Analysis	15
Pearson Correlation	10
MFDBSCAN-PC	**8**

The Table 1 shows the number of features selected by four different feature selection models; the proposed model MFDBSCAN-PC selects least number of attributes while BFS selects highest number of features.

Table 2: Performance comparison based on Accuracy

Feature Selection Models	Accuracy
Whole DS	90.58
BFS	72.78
PCA	75.23
PCS	80.64
MFDBSCAN-PCS	98.73

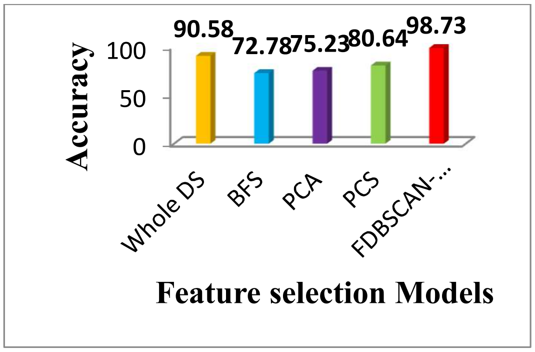

Figure 2: Performance comparison based on Accuracy

The table 2 and the figure 2 explore the performance of four different clustering models to predict CKD among diabetic patients. It shows that the proposed MFDBSCAN-PC produced least number of features with highest accuracy rate. This is because the proposed model handles the issue of vagueness and uncertainty of dealing with outliers in CKD dataset. The conventional feature selection models are not able to detect outliers in a prominent way. The DBSCAN with the moth flame optimization selects the parameter values prominently and overcome the problem of global searching.

Table 3: Performance comparison based on Precision

Feature Selection Models	Precision
Whole DS	0.914
BFS	0.728
PCA	0.764
PCS	0.814
FDBSCAN-PCS	0.983

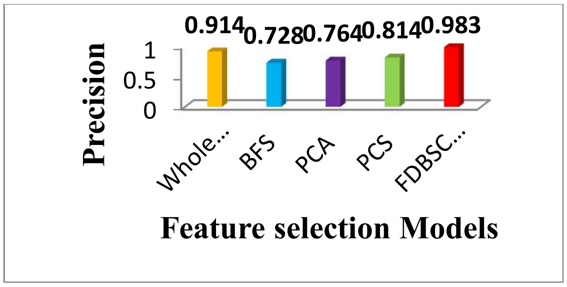

Figure 3: Performance comparison based on Precision

The table 3 and the figure 3 depict the performance of proposed feature selection models based on precision rate in CKD detection. The proposed model achieves better precision compared to other three models because similar instances are clustered using Moth flame optimization to discover the optimized value for the parameter fine tuning. Thus, it represents each instance in terms of finding correlation among the clustered features and overcomes the problem of outliers.

Table 4: Performance comparison based on Recall

Feature Selection Models	Recall
Whole DS	0.906
BFS	0.712
PCA	0.752
PCS	0.809
FDBSCAN-PCS	0.987

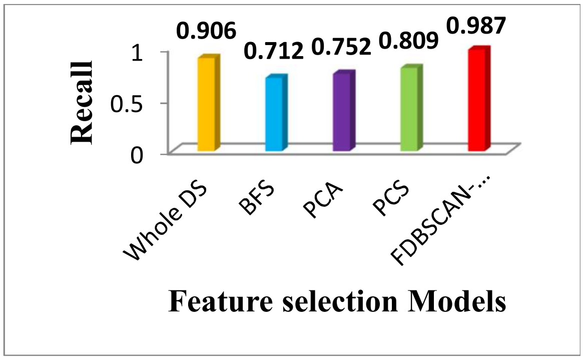

Figure 4: Performance comparison based on Recall

From the figure 4 it illustrates that the performance of proposed feature selection model MFDBSCAN-PC increased recall value in a better manner with reduces feature set. By reducing dependent variables, minimizing redundancy and maximizing the relevancy by finding correlation among the feature's during clustering process. The conventional feature selection model doesn't focus on removing irrelevant attributes in an imprecise dataset. Hence, the conventional feature selection model produces less recall value compared to the proposed model.

Table 5: Performance comparison based on FMeasure

Feature Selection Models	F-Measure
Whole DS	0.910
BFS	0.720
PCA	0.758
PCS	0.811
FDBSCAN-PCS	0.987

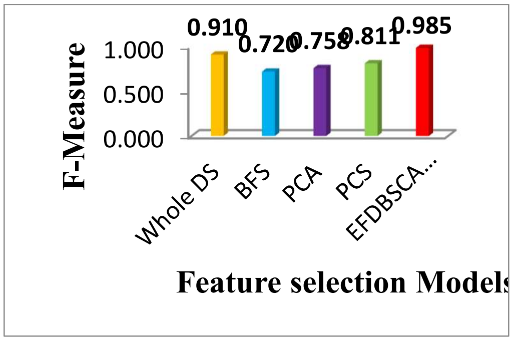

Figure 5: Performance comparison based on F-Measure

The F-measure value of four different feature selections models is displayed in the Table 5 and Figure 5 for chronic kidney disease detection with diabetic patients. This measure is influenced by both precision and recall henceforth it produce highest F-measure rate by proposed feature selection model. The proposed model initially discovers similarity among instances and clusters the similar instances as a group. In each group the influence of attributes is examined based on their significant score. The attributes with highest score are used as significant features to detect chronic kidney disease.

Table 6: Performance comparison based on Error Rate

Feature Selection Models	Error Rate
Whole DS	0.102
BFS	0.398
PCA	0.364
PCS	0.259
FDBSCAN-PCS	0.012

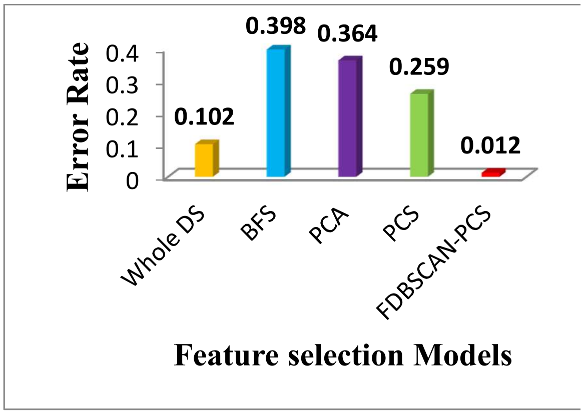

Figure 6: Performance comparison based on Error Rate

The Table 6 and the Figure 6 explore the error rate produced by each feature selection model by comparing the feature subset generated by four different clustering models BFS, PCA, PCS and proposed EFDBSCAN-PC, along with whole dataset for chronic kidney disease detection with diabetic patients. The error rate of the proposed EFDBSCAN-PC is very less compared to other models because the DBSCAN instead of assigning the parameter value's randomly, it uses the metaheuristic model, Moth Flame Optimization to search for the best fittest values using their spiral movement and flight path behavior and generates the optimized values to those parameter's. While searching for potential features, they are clustering, the relevancy among them is determined by applying Pearson correlation and the irrelevant attributes are eliminated. From the observed results it is proved that while providing enough knowledge in selection of centroids, the process of feature selection can be effectively done. While using standard feature selection algorithms they work on the ranking score or merit value of each attribute's information gain. When there is an uncertainty in selection of attributes these standard algorithms fail to focus on it.

5. CONCLUSION

In this paper, the main objective of the presented model MFDBSCAN-PC is that, to identify the potential attributes which improve the detection rate of chronic kidney disease in diabetic patients. This newly constructed model uses the bio inspirational behavior of the moth flame optimization to optimize the DBSCAN clustering performance. The conventional clustering models suffers from the local optima and earlier convergence, which search the significant features based on the distance measure. This work used the density-based feature clustering and the moth flame fitness value is used to assign the optimized value to the parameters of epsilon and minimum points used in DBSCAN. The relevancy among the attributes and the class variables are determined among each cluster using the Pearson correlation and the attributes which does not comply on any clusters is eliminated from the set. The simulation results are also proved that the error rate in chronic disease detection is greatly reduced after performing feature reduction using proposed MFDBSCAN-PC instead of using whole dataset. As an extension of this work in future, parallel implementation of moth flame during clustering process can be considered. Apart from classical theory-based feature selection, uncertainty theories can also be used for handling outlier in the CKD dataset.

REFERENCES:

[1] Arasu, S. D, Tirumalaiselvi, R. “A novel imputation method for effective prediction of coronary kidney disease”. *In 2017 2nd International Conference on Computing and Communications Technologies* (ICCCT), 127–136 (2017).

[2] BUPA Medical Research. UCI Machine Learning Repository, Available at: https://archive.ics.uci.edu/ml/datasets/liver+disorders, accessed April 2019.

[3] Chronic Liver Disease. Available at: https://stanfordhealthcare.org/medical-conditions/liverkidneys-and-urinary-system/chronic-liverdisease.html, accessed June 2019

[4] Gope, Sadhan, et al. "Moth Flame Optimization based optimal bidding strategy under transmission congestion in

deregulated power market." *Region 10 Conference (TENCON), IEEE,* 2016

[5]https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[6] K. Pavya B. Srinivasan, "Feature selection algorithms to improve thyroid disease diagnosis", 2017, *International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT)*, 2017, pp. 1-5.

[7] K. A. Padmanaban and G. Parthiban, "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease," *Indian Journal of Science and Technology*, vol. 9, (29), 2016

[8] Mohamed Elhoseny, K. Shankar, J. Uthayakumar, "Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease", *Scientific Reports*, (2019) 9:9583

[9] P Jinpon, M Jaroensutasinee and K Jaroensutasinee. "Risk assessment of type 2 diabetes mellitus in the population of Chonburi", Thailand. *Walailak J. Sci. Tech.* 2017; 14, 25-33.

[10] Ruiz-Arenas R, Sierra-Amor R, Seccombe D, R., Seccombe, D, Raymondo S, Graziani, M. S, Panteghini M, Adedeji T. A, Kamatham S. N, Biljak, "A Summary of Worldwide National Activities in Chronic Kidney Disease (CKD) Testing". *EJIFCC.* 2017;28(4):302-314. Published 2017 Dec 19.

[11] Rajathi, S, Radhamani, G. (2016, March). "Prediction and analysis of Rheumatic heart disease using KNN classification with ACO". *In Data Mining and Advanced Computing (SAPIENCE), International Conference* on (pp. 68-73). IEEE.

[12] Smita Chormungea, Sudarson Jenab, "Correlation based feature selection with clustering for high dimensional data", *Journal of Electrical Systems and Information Technology*, Volume 5, Issue 3, December 2018, Pages 542-549

[13] S. Naganjaneyulu, B.S. Rao, "A Novel Feature Selection Based Classification Algorithm for Real-Time Medical Disease Prediction", *IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing*, pp. 275-282, 2018

[14] Sultana, M., Haider, A., & Uddin, M. S. (2016, September). "Analysis of data mining techniques for heart disease prediction". *In Electrical Engineering and Information Communication Technology (ICEEICT), 2016 3rd International Conference on* (pp. 1-5). IEEE.

[15] Schubert, Erich Hess, Sibylle Morik, Katharina (2018). "The Relationship of DBSCAN to Matrix Factorization and Spectral Clustering". *Lernen, Wissen, Daten, Analyse (LWDA)*. pp. 330