Ensemble swarm behaviour based feature selection and support vector machine classifier for chronic kidney disease prediction

Article *in* International Journal of Engineering & Technology · May 2018

DOI: 10.14419/ijet.v7i2.31.13438

2 authors:

Belina Sara
SRM Institute of Science and Technology
12 PUBLICATIONS **144** CITATIONS

SEE PROFILE

Kalaiselvi K Dr
Saveetha College of Liberal Arts and Sciences
60 PUBLICATIONS **116** CITATIONS

SEE PROFILE

International Journal of Engineering & Technology, 7 (2.31) (2018) 190-195

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Ensemble swarm behaviour based feature selection and support vector machine classifier for chronic kidney disease prediction

S. Belina V.J. Sara[1*], K. Kalaiselvi [2]

[1]*Research Scholar, Department of Computer Science, School of Computing Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), (Formerly Vels University), Chennai.*

[2]*Associate Professor and Head, Department of Computer Science, School of Computing Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), (Formerly Vels university), Vels University, Chennai.*

**Corresponding author E-mail:belina_jyotsna@yahoo.co.in*

Abstract

Kidney Disease and kidney failure is the one of the complicated and challenging health issues regarding human health. Without having any symptoms few diseases are detected in later stages which results in dialysis. Advanced excavating technologies can always give various possibilities to deal with the situation by determining important realations and associations in drilling down health related data. The prediction accuracy of classification algorithms depends upon appropriate Feature Selection (FS) algorithms decrease the number of features from collection of data. FS is the procedure of choosing the most relevant features, removing irrelevant features. To identify the Chronic Kidney Disease (CKD), Hybrid Wrapper and Filter based FS (HWFFS) algorithm is proposed to reduce the dimension of CKD dataset. Filter based FS algorithm is performed based on the three major functions: Information Gain (IG), Correlation Based Feature Selection (CFS) and Consistency Based Subset Evaluation (CS) algorithms respectively. Wrapper based FS algorithm is performed based on the Enhanced Immune Clonal Selection (EICS) algorithm to choose most important features from the CKD dataset. The results from these FS algorithms are combined with new HWFFS algorithm using classification threshold value. Finally Support Vector Machine (SVM) based prediction algorithm be proposed in order to predict CKD and being evaluated on the MATLAB platform. The results demonstrated with the purpose of the SVM classifier by using HWFFS algorithm provides higher prediction rate in the diagnosis of CKD when compared to other classification algorithms.

Index terms: *Chronic Kidney Disease (CKD), Enhanced Immune Clonal Selection (EICS), filter, wrapper methods, Feature selection, prediction, Support Vector Machine (SVM), University of California Irvine (UCI).*

1. Introduction

Chronic Kidney Disease (CKD) is a universal problem with a moderately ever-increasing occurence, generality and of elevated consequences. About 11.2% of the adult people among overall world population are suffered from CKD [1], where in the USA ranges more than 27 million [2]. As per research carried out, 59% of each and every one of Americans are at risk stage of growing CKD in their natural life [3]. Increase of this problem is incompletely clarified by the growing occurrence of diabetes and hypertension which are considered to be highest risk factors for CKD. Medical diagnosis in this case varies accordingly to the different type of kidney related diseases, inconsistency in degree of development of disease and the exigent risk of heart related problems tends to mortality [4-5]. Correct detection of risk factor might help the individual in taking decision, enabling early and suitable patient care [6-7].

The target of medical diagnosis is to extract vital information from the huge medical datasets which are combined and put together often for all medical related issues. Broad studies have been done on cancer patients and various collection of medical datasets are used for medical diagnosis. "Drilling down of data" is described as the key idea for Knowledge Discovery in Databases (KDD). This procedure includes many methods with the purpose of suitable computational complexity [8]; KDD can also be defined as the process of selecting legitimate, original, potentially practical, and eventually comprehensible patterns in data. These approaches are like energy boosters and their area of applications have become increasingly indispensable for various health line concerned organizations. This helps to make healthy decisions based on the proper analysis techniques with vast amounts of medical set of data validated by healthcare connections. Mining data is being popular in the field of medical issues, even if it is not much needed, and many other factors initiated with the applications in field of medicine, such as detection, ability of response of data, and benefit of healthcare dealers [8]. Another purpose is that removal of unnecessary datas can benefit in healthy decision making by finding out the patterns and associations in immense amount of data [8-9]. Classification is considered to be a major spot in data which also recognized as supervised learning which refers to grouping. It is a method in which opinion and stuffs are acknowledged, renowned, differentiated, and unspoken. It predicts uncompromising category of labels and develops a model based on the instruction set and the class labels to sort new assortment of data.

If the class labels are commonly distributed most of the standard machine learning algorithms for data classification can be applied very efficiently for classification precision. Moreover, these standard algorithms show an average learning execution in case of classifying the huge data that have variation in the class labels. Since many inadequate attributes may be available in data to be extracted from massive dataset. Hence they need to be eliminated.

Most of the mining algorithms do not produce good outputs with large amounts of features or attributes. Hence Feature Selection (FS) techniques have to be applied before any of the machine learning algorithm is implemented, which highly minimizes the evaluation cost, restricts over fitting and upgrade the overview capacity [10-11].

FS has been classified into three major types [12]: filter, wrapper and embedded methods.

Filter [13] methods try in the direction of estimation of the significance of features with regard to heuristic valuing criteria in lacking of any specific classifiers. Commonly, it minimizes the computational complexity. Wrapper [14] method finds the space of feature, by making use of a classifier as the value for a candidate feature subset. This method performs based on the search strategy, and next to it, the classifier is skilled and experienced to finalize the applicant trait subset[15]. But it easily occurs computationally complexity problem. Moreover, they may yield feature subsets that are used as a classifier and are adequately overfitting.

An embedded method efficiently uses the construction of precise classes of learning classifiers in the direction of assist the feature selection procedure, and the significant module is derived from the basic information of a precise class of cataloging task. To obtain the accuracy of classification algorithms, one or more algorithms can be merged to achieve the reasonable accuracy. To diagnose the CKD, hybrid wrapper and filter based FS algorithm namely HWFFS is proposed in this work to minimize the dimension of CKD dataset. In wrapper based FS algorithm, the technique of the Enhanced Immune Clonal Selection (EICS) algorithm is followed to choose the vital features in the CKD dataset

2. Literature review

Kunwar et al [16] proposed a many classifiers such as Naive Bayes(NB) and Artificial Neural Network(ANN)to predict CKD. The experimental results which are applied in Rapid miner tool prove that NB produces more spot-on consequences than ANN. Moreover other interpretations can be carried out using other classifiers like Fuzzy logic, KNN algorithms which demonstrates that the MLP and C4.5 provides better results rather than other classifiers. Salekin and Stankovic [18] developed an asolution to find out CKD and explore 24 parameters which are basically related to kidney issues. Here the FS is also introduced to select optimal features for classifying CKD and rank them based on their accuracy. It provides good level of accuracy and at low cost.

Di Noia et al [19] proposed an ANN to classify health status in their patients potentially leading to End Stage of Kidney Disease (ESKD). Padmanaban and Parthiban[20] proposed a NB and Decision tree for prediction of Heart Disease dataset. From the results it can be demonstrate with the purpose of the accuracy is up to 91% for DT classification.

Potharaju and Sreedevi[22] discussed systematic way to address the imbalanced data classification problem by implementing the rule based ensemble learning techniques like bagging, boosting, voting and stacking to build models, also to develop the performance of learning algorithms. In this research, the importance is given to the preferred real data of CKD which is gathered from Apollo Hospitals, Tamil Nadu, and India, to predict kidney disease of patients. The obtained results prove that the model template which is selected to reduce the problem of misclassification of imbalanced data efficaciously. But this model template cannot classify accurately when imbalanced rate of class rises i.e. in case of massive data. For better result of imbalanced Big Data, new algorithmic plan of action has to be used which can be calculated using Hadoop framework and map reduce programming model.

An intellectual system expansion approach [23] has been proposed in our study. Performances were precisely validated in requisites of four vital arrangement assessment parameters. From the obtained results, more concentration is given minimized features for selecting CKD and thereby reducing improbability, decreases lapsing of time, and with higher accuracy.

3. Proposed methodology

In this study, the results of the SVM classifier is increased by using Hybrid Filter and Wrapper based Feature Selection (HFWFS). HFWFS have been used to minimize the dimension of features for the diagnosis of CKD. Filter methods uses a standard function to choose important features from the dataset. Wrapper based Feature Selection (WFS) algorithm adopts the procedure of Enhanced Immune Clonal Selection (EICS) algorithm. Datasets were obtained from University of California Irvine (UCI) machine learning repository. CKD consist of four major stages I–V with computed Glomerular Filtration Rate (GFR) shown in Table 1[24]. GFR is computed by using the parameters like serum creatinine, age, sex, body size, ethnic origin, etc.[24-26].

Table 1: The stages of CKDs

Stages	Clinical Features	GFR(mL/min/1.7 m^2)
I	Increased GFR	≥90
II	Damage with a mild decrease in GFR	60-89
III	Moderate Decrease in GFR	30-59
IV	Severe Decrease in GFR	15-29
V	Kidney Failure	<15 or dialysis

Table 2 represents the CKD data set from UCI that contains 24 attributes and additional one more attribute for class (binary) [26]. It contains 400 samples to two different classes ("CKD" ‑.250 cases; "NOTCKD" ‑.150 cases). Among the 24 attributes, 11 are numeric and 13 are nominal. The data set contains few missing values. Eliminating these tuples with missing values, 160 samples were used in this work.

Table 2: The attributes of CKD of UCI

Attribute number	Attributes	Attribute values	Attribute codes
1	Age	Years	Age
2	Blood pressure	mm/Hg	bp
3	Specific gravity	1.005, 1.010, 1.015, 1.020, 1.025	sg
4	Albumin	0, 1, 2, 3, 4, 5	al
5	Sugar	0, 1, 2, 3, 4, 5	su
6	Red blood cells	Normal, abnormal	rbc
7	Pus cell	Normal, abnormal	pc
8	Pus cell clumps	Present, not present	pcc
9	Bacteria	Present, not present	ba
10	Blood glucose random	mg/dl	bgr
11	Blood urea	mg/dl	bu
12	Serum creatinine	mg/dl	sc
13	Sodium	mEq/L	sod
14	Potassium	mEq/L	pot
15	Hemoglobin	g	hemo
16	Packed cell volume	-	pcv
17	White blood cell count	cells/cumm	wbcc
18	Red blood cell count	millions/cmm	rbcc
19	Hypertension	No, yes	htn
20	Diabetes mellitus	No, yes	dm
21	Coronary artery disease	No, yes	cad
22	Appetite	Good, poor	appet
23	Pedal edema	Yes, no	pe
24	Anemia	Yes, no	ane
25	Class	CKD, NOTCKD	-

1. Filter methods

Filter method selects the features whose ranks are the higher than other features, and then the chosen subset features can be utilized for any predication algorithm. The following three type of filter function are used in this work.

Information Gain(IG)

Information gain is a measure of the variation among two probability distributions. It calculates a feature X by computing the level of information gained with regard to the class variable Y, described as follows:

$$I(X) = H\left(P(Y) - H\left(P\left(\frac{Y}{X}\right)\right)\right) \quad (1)$$

If X is not differentially expressed, Y will be independent of X, thus X will have small information gain value, and vice versa [27].

Correlation Based Feature Selection (CFS)

Correlation based Feature Selection (CFS) is a filter algorithm with the intention of grades feature subsets. CFS's feature subset evaluation function is shown as follows [28]:

$$Merits_S = \frac{kr_{Cf}}{\sqrt{k + (k+1)r_{ff}}} \quad (2)$$

where Merits is the heuristic "merit" of a feature subset S consisting of k features, rcf is the mean feature-class correlation , and rff is the average feature-feature inter-correlation. The heuristic handles irrelevant features as they will be bad predictors of the class. Redundant attributes are differentiated as they will be highly correlated with many features.

Consistency Based Subset Evaluation (CS)

CS follows the class consistency rate as per the estimation measure. The objective design is to obtain a set of features with the purpose of separate the new dataset into subsets which include one class majority [29]. One of the popular CS is consistency metric is described as follows:

$$Consistency_s = 1 - \frac{\sum_{j=0}^{k} |D_j| - |M_j|}{N} \quad (3)$$

where s is feature subset, D_j is the number of rate of the j^{th} features value combination, M_j is the cardinality of the best part class for the j^{th}feature value, and N is the total number of features in the dataset [29].

2. Wrapper methods

Wrapper method evaluates the scores of feature sets with the purpose of depend on the predictable power with the help of a classifier algorithm as a black box. This feature gains the space of each and every feature of subsets.

Clonal Selection Algorithms (CSAs)

The major objective of CSA theory is in the event where B cell acts to invaded antigen through modifying the receptor called antibody is illustrated in figure 1. In general CLOGNALG [30], is one of the description for CSAs. Three major operations are cloning, hypermutation, and selection is used for choosing most the features in CKD dataset samples. To overcome the prediction error of the SVM classifier in CKD and the complexity of the coding, Enhanced Immune Clonal Selection (EICS) algorithm is coded in real number and each CKD feature dimension of an attribute reduction is considered as an attributes segment. According to the recombination in immunology, any orderly rearrangement of features segments would establish a new B cell.

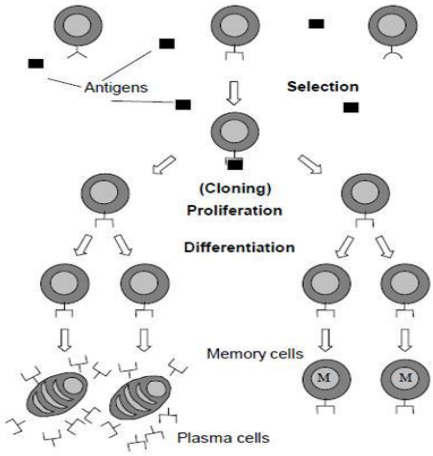

Figure 1: Clonal selection principle

As shown in Figure 2, the recombination could be (a) between two CKD samples as crossover or (b) between many features as the combination of randomly selected CKD features. With the use of normalization, the combination of CKD features will be in the determined order, such as in Figure 3(a) or could be randomly arranged as shown in Figure 3(b) in the computational respective,

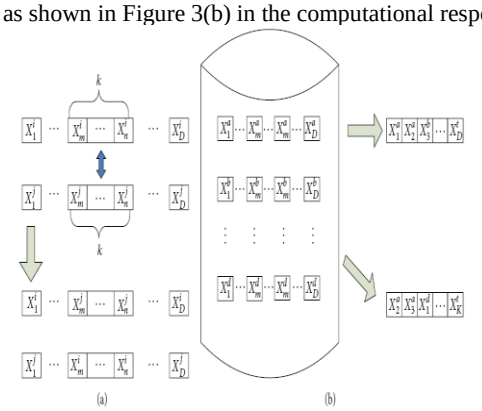

Figure 2: The way of rearrangement of gene segments (a) between two CKD dataset samples (b) among several CKD samples

Select two dataset samples of features in random from the CKD dataset, denoted by X_A , X_B, and then randomly select m number of features $m \in [1,D]$, from each of them, in which feature index could be recorded as vectors FV_A and FV_B, respectively. The new selected features samples are created by

$$X_A^{'FV_A} = \alpha X_A^{FV_A} + (1-\alpha) X_B^{FV_B} \quad (4)$$

$$X_B^{'FV_B} = \alpha X_B^{FV_B} + (1-\alpha) X_A^{FV_A} \quad (5)$$

where α is a randomly selected number between 0 and 1.It should be noted that the range of each CKD features of decision variable should be normalized at first. The new EICS algorithm the recombination operator is represented in Figure 3. New features are generated through the combinational recombination which is illustrated by Figure 4. In the example, m equals 3. It requires to be known that i_1 could be different from j_1, the same as in i_2 and j_2 and i_3 and j_3 as long as the normalization has been done.

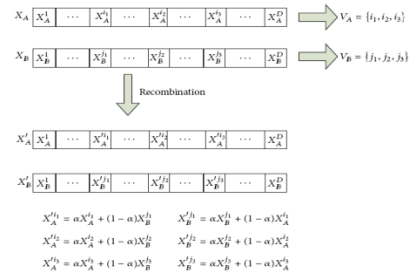

Figure 3: Recombination process of EICS

Evaluation is done on the fitness of the new selected features of CKD samples. Together with the original CKD samples, two with the higher fitness will survive, and the other two CKD samples of features are eliminated; Select features with high fitness from the

CKD dataset. Hypermutation operator brings diversity for the population by introducing perturbation for each clone. Although there exists many methods to implement this operator [31], alternatively proportional strategy remains to be in main basis. The concept of the operator [32] is followed in this work, where every solution is subject to M mutations. The inversely proportional law is used to evaluate the number of the mutations M:

$$\alpha = \exp\left(-\rho f^*(X_i)\right) \quad (6)$$

$$M = \lfloor(\alpha \times n) + 1\rfloor \quad (7)$$

where $f^*(X_i) \in [0,1]$ is the normalized fitness of X_i, ρ is the decay constant which evaluates the nature of the mutation rate, and $\lfloor\cdot\rfloor$ returns the lower bound integer.

$$X_i^{'j} = \begin{cases} X_{r1}^j + \lambda\left(X_{r1}^j - X_{r2}^j\right) if j \in randM(n) \\ X_i^j elsewhere \end{cases} \quad (8)$$

Then, M mutation is done on each candidate solution: (j) is the j^{th}feature of the i^{th} dataset, rand (n) $\in \{1, \ldots, n\}$ is randomly chosen M indexes without repetition, and $\lambda \in [-1, 1]$. $r1, r2 \in \{1, 2, \ldots,\}$ are randomly chosen numbers; The amplitude of the hypermutation is determined automatically by the variation of randomly chosen features in the dataset. The mutation equation (8) could be taken as the variant of differential evolution The M strategy determines the direction including the number of dimensions, while the equation determines the distance of the mutated clones with their parents. With union of both, the amplitude of the hypermutation is automatically determined with regard to the distribution of the population is shown in algorithm 1.

Algorithm 1: Proposed EICS algorithm

1. Initialization of the CKD with D number of features $X_i = (X_i(1), \ldots X_i(D))$, $i = 1, \ldots N$ be the number of dataset samples , $j = 1$ to D be the number of features is produced randomly within the range of boundaries of the decision space:
   $$X_i^j = X_{\min}^j + rand(0,1)\left(X_{\max}^j - X_{\min}^j\right)$$
   $X_{\min}^j$&$X_{\max}^j$ are the lower and upper bound value of features j respectively .
2. Evaluation classification accuracyof the antibody population as their fitness
3. Generate new copies of the features as antibodies. The recombination rate is set to N_r.
4. Mutate all the generated copies(hypermutation)
   (i) Cloning: each CKD samples X_i generates N_c copies $\{X_i^1, X_i^2, \ldots, X_i^{N_c}\}$, where $N_c$is the clone number, which is a user defined constant
   (ii) Hypermutation: each clone $X_i^j, j = 1, \ldots, N_c$, goes through the hypermutation and generates the hypermutated clones $X_i^{'j}$.
   (iii) Selection: select the CKD sample of features with highest fitness among X_i and hypermutated clones $\{X_i^{'1}, X_i^{'2}, \ldots, X_i^{'N_c}\}$.
5. Select the features with highest classification accuracy to survive(Selection)
6. Repeat Steps 2–5 until a it reaches termination criterion is met

3. Support Vector Machine (SVM) classifier

SVM has a better potential in classification problems. It be able to extend the generalization results by dealing with mapping the inputs into high dimensional areas and evaluating the of quadratic programming classification problem. It is able to locate the optimum disjunctive hyperplane. By considering training samples (Eq. 9), each disjunctive hyperplane must prepare two constraints (Eq. 10) for two classes.

$$(x_i, y_i) | x_i \in R^N, y_i \in \{-1,1\}, i = 1, .. n, \{(w, x_i) + b \gg +1 - \varepsilon_i\}, if y_i = +1 \quad (9)$$

$$(w, x_i) + b \gg -1 + \varepsilon_i if y_i = -1 \quad (10)$$

Inequalities of Eq. (10) are equal to Eq. (11). Minimizing Eq(12) to Eq(11) can progress the hyperplane separation.

$$y_i[(w, x_i) + b] \gg +1 - \varepsilon_i, i = 1, \ldots n \quad (11)$$

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{n} \varepsilon_i \quad (12)$$

In Eq(12) ,C parameter can determine the stability among complication and correctness of classifier. In respect to this, Lagrange multipliers be able to determine the solution for convex optimization problem and by using appropriate substitution it be able to achieve the optimized solution for Eq(12).Decision function will be given in Eq(14) by Lagrange multiplies given in Eq(13) [33].

$$Maximize: \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i, x_j) \quad (13)$$

$$Subjectto: \sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \gg 0, \forall i, f(x) = sgn\left(\sum_{i=1}^{n} \alpha_i y_i (x_i, x) + b\right) \quad (14)$$

The classifier results are experimented using the MATLAB tool.

4. Experimental results

The classification algorithms and feature selection algorithms were implemented in MATLAB tool. Used Intel Core 2 Duo Processor E7400 CPU (2.8 GHz Dual Core, 1066 MHz FSB, 3 MB L2 cache) with 2 GB RAM for implementation. The following metrics have been used and discussed as follows. In order to predict the performance of the system, computed Classification Accuracy (CA), specificity, sensitivity, precision, F-measure and Error Rate (ER) as these are very vital parameters to predict the performance of the system without the knowledge of distribution of data. Computed True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) to further compute other performance parameters

Accuracy is defined as the ability of classifier algorithm to diagnose of classes of dataset

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (15)$$

Sensitivity is defined as the accuracy measure of the target class's occurrence (Eq 16).

$$Recall(Re) = Sensitivity = \frac{TP}{TP + FN} \times 100 \quad (16)$$

Specificity relates to the test's capability to correctly detect patients without a condition.

$$Specificity = \frac{TN}{TN + FP} \times 100 \quad (17)$$

Precision is defined as the division of relevant instances between the correct instances

$$Precision(Pr) = \frac{TP}{TP + FP} \times 100 \quad (18)$$

F-measure is defined as the harmonic mean of precision and recall is measured as follows:

$$F - measure = 2 \times \frac{Re \times Pr}{Re + Pr} \times 100 \quad (19)$$

Figure 5 shows the performance comparison results of the sensitivity and specificity metrics with respect to three different classifiers such as the NB, ANN and SVM. However the proposed SVM-HWFFS algorithm produces higher sensitivity results of 90.91% which is 6.91% and 9.09% higher when compared to ANN-HWFFS and NB-HWFFS algorithm. The other classifiers such as NB-HWFFS and ANN-HWFFS classifiers produces sensitivity results of 81.82% and 84.00 % respectively. Similarly proposed SVM-HWFFS algorithm produces higher specificity results of 87.5% which is higher than other two classifiers. The other classifiers such as NB-HWFFS algorithm and ANN-HWFFS

classifier produces specificity results of 62.50% and 62.50% respectively. The values of all the classifiers are discussed in table 4.

Figure 6 shows the results of the precision and recall metrics with respect to three different classifiers such as the NB, ANN and SVM. However the proposed SVM-HWFFS algorithm produces higher precision results of 95.24% which is 7.74% and 9.53% higher when compared to ANN-HWFFS and NB -HWFFS algorithm. The other classifiers such as ANN-HWFFS and NB-HWFFS algorithm produce precision results of 85.71% and 87.50% respectively. The values of all the classifiers are discussed in table 4.

Similarly proposed SVM-HWFFS algorithm produces higher f-measure results of 93.02% which is 7.31% and 9.3% higher when compared to ANN-HWFFS and NB-HWFFS classifiers respectively. The other classifiers such as NB-HWFFS algorithm and ANN-HWFFS classifier produces f-measure results of 83.72% and 85.71% respectively. Figure 7 shows the results of the accuracy and error rate metrics with respect to three different classifiers such as the NB, ANN and SVM. However the proposed SVM-HWFFS algorithm produces higher accuracy results of 90% which is 11.21% and 13.33% higher when compared to ANN-HWFFS and NB -HWFFS algorithm produces lesser error rate results of 10% which is 11.21% and 13.33% lesser when compared to ANN-HWFFS and NB-HWFFS classifiers respectively. The other classifiers such as NB-HWFFS algorithm and ANN-HWFFS and algorithm produces accuracy results of 78.79% and 76.67% methods respectively. The values of all the classifiers are discussed in table 4. The True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) of the SVM-HWFFS classifier are 100, 35, 05 and 10 respectively. These parameter values of all the classifiers are discussed in table 3.

Table 3: Confusion Matrix vs. Classifiers

Class	NB		ANN		SVM		NB-HWFFS		ANN-HWFFS		SVM-HWFFS	
	True positive	True negative	True positive	True negative	True positive	True negative	True positive	True negative	True positive	True negative	True positive	True negative
Predictive positive	90	20	80	20	85	15	90	15	105	15	100	05
Predictive negative	30	10	25	25	25	25	20	25	20	25	10	35

Table 4: Performance Metrics vs. Classifiers

Methods	Results(%)					
	Sensitivity	Specificity	recision	F-measure	Accuracy	Error rate
NB-HWFFS	81.82	62.50	85.71	83.72	76.67	23.33
ANN-HWFFS	84.00	62.50	7.50	85.71	78.79	21.21
SVM-HWFFS	90.91	87.50	5.24	93.02	90.00	10.00
NB	75.00	33.33	1.82	78.26	66.67	33.33
ANN	76.19	55.56	0.00	78.05	70.00	30.00
SVM	77.27	62.50	5.00	80.95	73.33	26.67

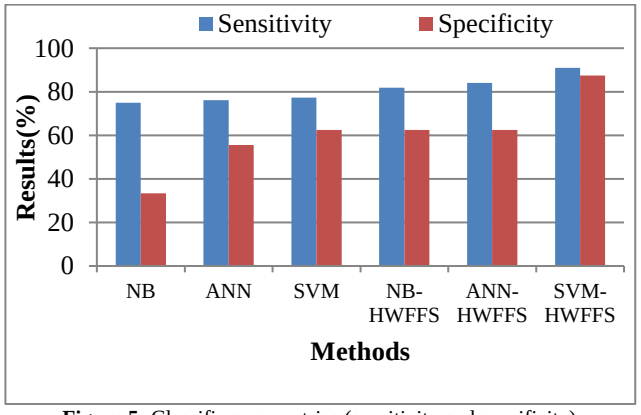

Figure 5: Classifiers vs. metrics (sensitivity and specificity)

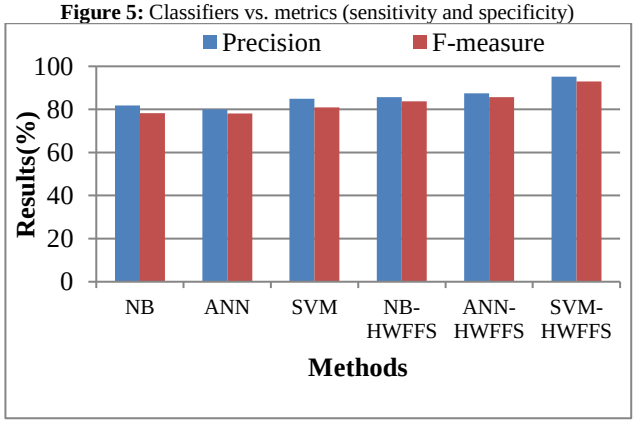

Figure 6: Classifiers vs. metrics (Precision and F-measure)

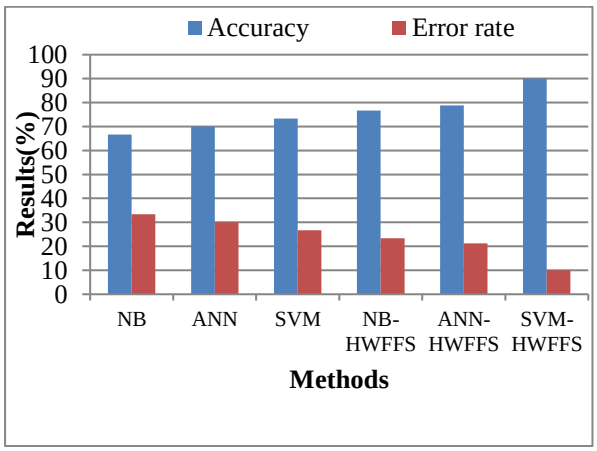

Figure 7: Classifiers vs. metrics (Accuracy and Error rate)

5. Conclusion and future work

In our paper we have introduced a original Hybrid Wrapper and Filter based FS (HWFFS) algorithm to choose optimal subset of features from the datasets to predict CKD with Support Vector Machine (SVM) classifier. According to the UCI statistics set, there are 24 features for predicting CKD or non-CKD. At least there are 16 features are selected from original dataset. The major objective this work is to predict CKD or non-CKD with reasonable correctness using selected features from HWFFS algorithm. An intellectual HWFFS approach adopts the procedure of filter and wrapper algorithm which has been used in this study. However the wrapper based FS algorithm is performed based on the Enhanced Immune Clonal Selection (EICS) to find out the selected features that describe the data set much better, thus increases the accuracy of the classifier. The popularly known SVM classifier have been followed for the strength of the condensed feature set. Results were determined in terms of six chief cataloging metrics. The resutls concludes that the proposed SVM with HWFFS algorithm yields lesser error rate results of 10.00% ,whereas other methods such as NB and ANN produces 23.33% and 21.21%. Hence it concludes that the proposed work performs better than other classifiers. As demonstrated from the results, concentration is more on the reduced features which is used for specifying CKD and by this means reducing ambiguity,

decreasing usage of time, and reduces costs. This dataset contains a number of noisy and missing values. Therefore, a new classifier is required to dealing with missing and noisy values are kept as future work. Associating the data mining results by the expert systems prove in the direction of indicate the factors having the highest risks on CKD. The guidelines specified by the data mining methods include association rules form to describe the function of CKD factors.

References

[1] Zhang QL & Rothenbacher D, "Prevalence of chronic kidney disease in population-based studies: systematic review", *BMC public health*, Vol.8, No.1, (2008).

[2] Baumgarten M & Gehr T, "Chronic kidney disease: detection and evaluation", *American family physician*, Vol. 84, No.10, (2011).

[3] Moyer VA, "Screening for chronic kidney disease: Us preventive services task force recommendation statement", *Annals of internal medicine*, Vol.157, No.8, (2012), pp.567–570.

[4] Keane WF, Zhang Z & Lyle PA, "Risk scores for predicting outcomes in patients with type 2 diabetes and nephropathy: the RENAAL study", *Clin J Am SocNephrol.*, Vol.1, No.4, (2006), pp.761-767.

[5] Keith DS, Nichols GA & Gullion CM, "Longitudinal follow-up and outcomes among a population with chronic kidney disease in a large managed care organization", *Arch Intern Med.,* Vol.164, No.6, (2004), pp.659-663.

[6] Taal MW & Brenner BM, "Predicting initiation and progression of chronic kidney disease: developing renal risk scores", *Kidney Int.*, Vol.70, No.10, (2006), 1694-1705.

[7] Taal MW & Brenner BM, "Renal risk scores: progress and prospects", *Kidney Int.*, Vol.73, No.11, (2008), pp.1216-1219.

[8] Koh HC & Tan G, "Data mining applications in healthcare", *Journal of Healthcare Information Management*, Vol.19, No.2, (2005), pp.64–72.

[9] Han J & Kamber M, *Data mining: concepts and techniques*, 3rd ed. Burlington, MA: Elsevier, (2011).

[10] Liu Y & Zheng YF, "FS-SFS:A novel feature selection method for support vector machines", *Pattern Recognit.,* Vol.39, (2006), pp.1333–1345.

[11] Reunanen J, "Overfitting in making comparisons between variable selection method", *J. Mach. Learn. Res.,* Vol.3, (2003), pp.1371–1382.

[12] Brown G, Pocock A, Zhao MJ & Luján M, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection", *J. Mach. Learn. Res.*, Vol.13, No.1, (2012), pp.27–66.

[13] Kohavi R & John GH, "Wrappers for feature subset selection", *Artif. Intell.*, Vol.97, (1997), pp.273–324.

[14] Chen G & Chen J, "A novel wrapper method for feature selection and its applications", *Neuro computing*, Vol.159, (2015), pp.219–226.

[15] Uncu O & Turksen IB, "A novel feature selection approach: combining feature wrappers and filters", *Inf. Sci.*, Vol.177, (2007), pp.449–466.

[16] Kunwar V, Chandel K, Sabitha AS & Bansal A, "Chronic Kidney Disease analysis using data mining classification techniques", *6th International Conference Cloud System and Big Data Engineering (Confluence)*, (2016), pp.300-305.

[17] Boukenze B, Haqiq A & Mousannif H, "Predicting Chronic Kidney Failure Disease Using Data Mining Techniques", *Advances in Ubiquitous Networking*, (2017), pp.701-712.

[18] Salekin A & Stankovic J, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes", *IEEE International Conference on Healthcare Informatics (ICHI)*, (2016), pp.262-270.

[19] Di Noia T, Ostuni VC, Pesce F, Binetti G, Naso D, Schena FP & Di Sciascio E, "An end stage kidney disease predictor based on an artificial neural networks ensemble", *Expert Systems with Applications*, Vol.40, No.11, (2013), pp.4438-4445.

[20] Padmanaban KA & Parthiban G, "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease", *Indian Journal of Science and Technology*, Vol.9, No.29, (2016), pp.1-6.

[21] Mohammed Siyad B, Manoj M, Mohammed Siyad B & Manoj M, "Fused features classification for the effective prediction of chronic kidney disease", *International Journal*, Vol.2, (2016), pp.44-48.

[22] Potharaju SP & Sreedevi M, "Ensembled Rule Based Classification Algorithms for predicting Imbalanced Kidney Disease Data", *Journal of Engineering Science and Technology Review*, Vol.9, No.5, (2016), pp.201-207.

[23] Misir R, Mitra M & Samanta RK, "A reduced set of features for chronic kidney disease prediction", *Journal of Pathology Informatics*, Vol.8, No.1, (2017), pp.24-29.

[24] Levey AS & Coresh J, "Chronic kidney disease", *The Lancet*, Vol.379, (2012), pp.165-180.

[25] Levey AS, Coresh J, Bolton K, Culleton B, Harvey KS, Ikizler TA & Levin A, "K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification", *American Journal of Kidney Diseases*, Vol.39, (2002).

[26] Anderson J & Glynn LG, "Definition of chronic kidney disease and measurement of kidney function in original research papers: A review of the literature", *Nephrol Dial Transplant*, Vol.26, (2011), pp.2793‑2798.

[27] Lee IH, Lushington GH & Visvanathan M, "A filter-based feature selection approach for identifying potential biomarkers for lung cancer", *Journal of Clinical Bioinformatics*, Vol.1, No.1, (2011), pp.1-8.

[28] Hall MA, "Correlation-based feature selection for machine learning", PhD, Department of Computer Science, The University of Waikato, Hamilton, (1999).

[29] Arauzo-Azofra A, Benitez JM & Castro JL, "Consistency measures for feature selection", *Journal of Intelligent Information Systems*, Vol.30, No.3, (2008), pp.273-292.

[30] Wang Q, Wang C & Gao XZ, "A hybrid optimization algorithm based on clonal selection principle and particle swarm intelligence", *Sixth International Conference on Intelligent Systems Design and Applications*, Vol.2, (2006), pp.975-979.

[31] Jansen T & Zarges C, "Analyzing different variants of immune inspired somatic contiguous hyper mutations", *Theoretical Computer Science*, Vol.412, No.6, (2011), pp.517–533.

[32] Pavone M, Narzisi G & Nicosia G, "Clonal selection: an immunological algorithm for global optimization over continuous spaces", *Journal of Global Optimization*, Vol.53, No.4, (2012), pp.769–808.

[33] Xie J & Wang C, "Using support vector machines with a novel hybrid feature selection method for diagnosis of erythematosquamous diseases", *Expert Syst. Appl.,* Vol.38, (2011), pp.5809–5815.