A Hybrid Filter Wrapper Embedded-Based Feature Selection for Selecting Important Attributes and Prediction of Chronic Kidney Disease

K. Kalaiselvi and S. Belina V. J. Sara

Abstract Today's most significant healthcare problem that is prevailing is the chronic kidney disease (CKD). The disease integrates well-defined pathophysiological process that will be experimental for determining irregular kidney functions and the glomerular filtration rates. To forecast the disease, different data mining techniques are used to discover the connections between various elements, which can be utilized to determine the progress and status of CKD. Data is obtained from the patient's healthcare records. The main purpose of this research is to avail the Hybrid Filter Wrapper Embedded-Based Feature Selection (HFWE-FS), which will be utilized to select CKD datasets from potential feature subsets. HFWE-FS algorithm integrates the process of filtering, wrapping and embedding algorithms. The filter algorithms are integrated with reference on certain metrics: Gini index, gain ratio, One R and Relief. The wrapper algorithms via enhanced bat algorithms are purposed to select the analytical features from wide-range CKD sets of data. The embedded algorithms are underpinned, and this depends on the support vector machine (SVM)-t statistic, which selects the analytical features out of the wide-range CKD dataset. The results of the feature selection algorithms are integrated and identified as the HFWE-FS algorithm. The SVM algorithm for the CKD prediction is proposed as a final stage. The database used is taken from 'CKD' implemented on the MATLAB. The results perceived that the SVM classifier along with HFWE algorithm gets high classification rate when contrasted with other categorization algorithms: Naïve Bayes (NB), artificial neural networks (ANNs) and support vector machine (SVM) in CKD completion.

Keywords Chronic kidney disease (CKD) · Improved bat algorithm (IBA) · Feature selection (FS) · Hybrid Filter Wrapper Embedded (HFWE) · Classification · Support vector machine (SVM)

K. Kalaiselvi · S. B. V. J. Sara (✉)
Department of Computer Science, School of Computing Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India

A. Ramu et al. (eds.), *International Conference on Computing, Communication, Electrical and Biomedical Systems*, EAI/Springer Innovations in Communication and Computing, https://doi.org/10.1007/978-3-030-86165-0_14

1 Introduction

Chronic Kidney Disease (CKD) is progressively development in generally following months. In common it shouldn't be identified previously to it dropped to 25% of its functionality. The persons who get treatment for this shouldn't be affected and identified by renal failure because kidney failure couldn't provide some symptoms originally.

In accordance with the National Health Service, kidney disease is found to be predominant in Africa and South Asia compared to other countries. There is a need for early detection of kidney failure, through which both kidneys be capable under control and consequently mitigate the threat of irreversible issues [1]. CKD can be able to be identified through a blood test which distinguishes measuring factors, and thereby doctors can decide the treatment process that reduces the progression rate (Kathuria and Wedro [2]).

Filter, wrapper and embedded methods are the taxonomy of feature selection methods. The reavailability of the embedded algorithm and the filter wrappers for FS and the algorithm used for classification called SVM to characterize the subset which is chosen is illustrated in Fig. 1.

The wrapper method measures the feature sets that depend on the predictable influence by means of a classifier like a black box (Ladha and Deepa [3]). Classification system makes use of varied classifiers to reduce the features and classifiers to identify quite a few types of diseases [4].

The accuracy rate of these methodologies using feature selection is investigated in this study. For the diagnosis of CKD, Hybrid Filter Wrapper Embedded-Based FS algorithm is used to reduce the element of features, and consequently the features are classified using SVM. The collection of data was done from the University of California Irvine (UCI) repository. The element is diagnosed as given: earlier, in Sect. 2, a background analysis of the techniques utilized in feature selection

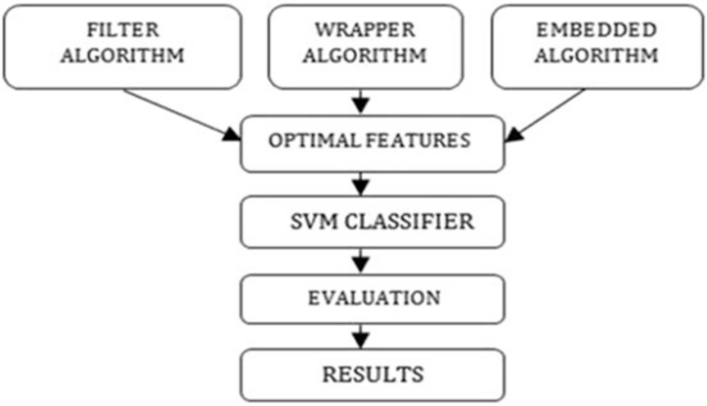

Fig. 1 SVM classification with Hybrid Filter Wrapper Embedded (HFWE) algorithm

purposed for further evaluation of the illness is conducted, and Sect. 3 details the methodology which is proposed. Section 4 shows the simulation results for the used methods and its comparison. In conclusion, Section 5 wraps up the paper.

2 Literature Review

[5] projected a system for forecasting the renal failure timeframe and named as adaptive neuro-fuzzy inference system (ANFIS) which used CKD which depends on live time clinical data. On comparing real data with predicted values, the study exposed to smooth the progress of the ANFIS model is capable of precisely estimating GFR variations in each and every sequential epoch and at long future periods in spite of elevated qualms of the human body and the energetic environment of CKD sequence.

[6] developed a method, which is too risky to estimate an individual's absolute risk of incident CKD. The participants were observed for a decade to assess the development of CKD. Performance measure evaluation was carried out using calibration and discrimination measures. Further investigation was proposed for the efficacy of this score in recognizing individuals in the community at risky chronic kidney illness.

[7] introduced a machine learning wrapper method for the identification of a set of 12 attributes which exhibits CKD detection with high accuracy. The experiment was conducted on a 400 individual's dataset, out of which 250 were detected for CKD. The results revealed that according to F1 measure the precision of 0.993 with 0.1084 root-mean-square error was attained.

[8] designed a technique that incorporates case-based reasoning (CBR) and data mining (DM) for forecasting and identifying chronic disease. These conclusions elucidate the references for the doctors as well as the patients who are into the CKD.

[9] IgA nephropathy (IgAN) is a wide-reaching disease that has an effect on human kidneys and shows the way to the end-stage kidney disease (ESKD), which typically necessitates renal spare therapy with kidney transplant or dialysis. Introducing an artificial neural network is to categorize patients' health prominence impending to ESKD. The developed tool is accessible together as a live web application and as an Android mobile application.

[10] conducted a study on the warning signs that are connected with the progressive development of CKD, with different frameworks, i.e. support vector machine (SVM), soft independent modelling of class analogy (SIMCA) and the k-nearest neighbour (KNN). The techniques are applicable in assessing the clinical data used from UCI machine learning repository.

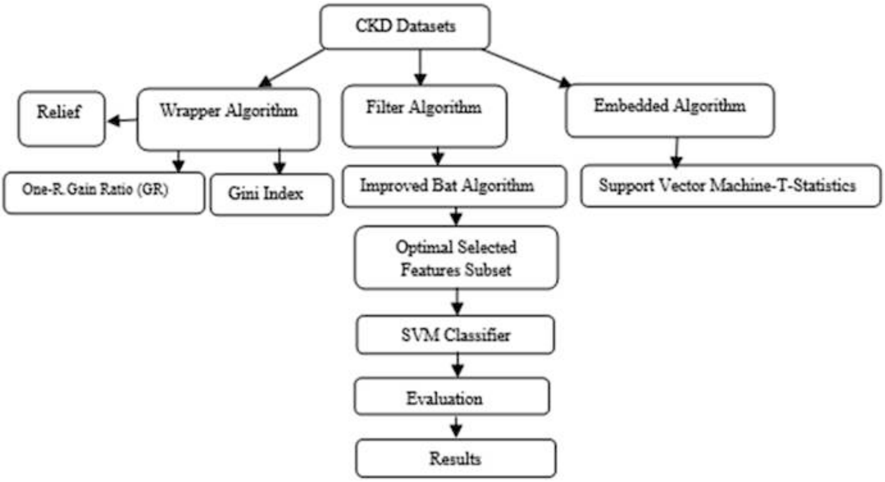

Fig. 2 Block diagram reavailable HFWE-FS

3 Proposed Methodology

For chronic kidney disease (CKD) development, early recognition and successful dealing are the simple treatment to decrease the death rate. The machine learning algorithm with respect to SVM was utilized in the direction of CKD forecasts in this work. In the prediction process of CKD, filter wrapper embedded techniques are utilized to potentially minimize the feature number in CKD set of data. In the technique, Gini index (GI), gain ratio (GR), One R, Relief and filter were utilized. In the wrapper technique, improved bat algorithm (IBA) was utilized. The embedded algorithm is typically executed based on the SVM-t statistics to select the probably analytical feature from wide-range CKD dataset from UCI repository for machine learning. Figure 2 shows the structure of the Hybrid Filter Wrapper Embedded (HFWE) for predictive attributes and SVM algorithm for the identification of the CKD.

3.1 Dataset Information

The rate of glomerular filtration (GFR) includes mathematical functions utilizing serum, origin, body size and age, among others. If a kidney's work is regular and polluted to scope, to make the blood unwell, unnecessary items are put together up to higher levels. A set of UCI CKD data contain 24 attributes and one additional (binary) [11] class attribute with 400 samples ("CKD"-.250 cases; "NOTCKD"-.150 cases).

3.2 *Wrapper Algorithm*

The achievement of feature sets is evaluated using a classifier.

3.2.1 Relief-F

An attribute selection technique called Relief-F determines features by how fine its value differentiates examples with the purpose are from several groups, however, are comparable to each other. For each feature Fr, Relief-F chooses a random sample and *k* of its nearest neighbours from the similar class and every class. Then *f* is scored as the sum of weighted variation in several classes and the similar class. If Fr is expressed in a differential form, its determination shows higher variation, for instance, from several classes, consequently, it will obtain a higher score [12]:

$$\mathrm{SC}_{\mathrm{R}}\left(\mathrm{fr}_i\right)=\frac{1}{p}\sum_{t=1}^{p}\left\{-\frac{1}{m_{\mathrm{fr}_{\mathrm{t}}}}\sum_{\mathrm{fr}_j\in\mathrm{NH}_{\mathrm{fr}_{\mathrm{t}}}}d\left(f_{t,i}-f_{j,i}\right)+\sum_{y\neq \mathrm{y}_{\mathrm{fr}_{\mathrm{t}}}}\frac{1}{m_{\mathrm{fr}_{\mathrm{t}}}}\frac{P(y)}{1-P\left(y_{\mathrm{x}_{\mathrm{t}}}\right)}\sum_{\mathrm{fr}_j\in NM(\mathrm{fr}_{\mathrm{t}},y)}d\left(f_{t,i}-f_{j,i}\right)\right\} \tag{1}$$

where $\mathrm{yfr}_{\mathrm{t}}$ is the sample class label, frt, and *P*(*y*) is the likelihood of a classy sample. NH(Fr) or NM(Fr,y) are the few points closest to the sample of Fr with a related class of Fr or a separate class (class *y*), respectively. The sizes of the $\mathrm{NH}(\mathrm{fr}_{\mathrm{t}})$ and $\mathrm{NM}(\mathrm{fr}_{\mathrm{t}}, y)$ sets are correspondingly $\mathrm{mfr}_{\mathrm{t}}$ and mfrt, respectively. The magnitude of both NH(Fr) and NM(Fr, *y*); ¥y x $\mathrm{yfr}_{\mathrm{t}}$ is typically set to an unvarying value *k*, which is a user-defined identifier.

3.2.2 One-R

One-R is an easy algorithm, and it considers each and every numerical feature as unbroken and makes use of a simple algorithm to separate the range of values addicted to a number of disjoint intervals. The gain ratio is a ratio of the information obtained to the intrinsic drilled down data and also unique to the equation of *y* means (2):

$$GR=\frac{IG}{\mathrm{H}(fr)} \tag{2}$$

As shown in Eq. (2), as soon as feature *Y* needs to be forecasted, the information gain (IG) has to be normalized by separating entropy of feature fr and likewise to the reverse also. Suitable in the direction of the normalization, the gain ratio ideals constantly decrease in the series [0,1]. The importance of gain ratio = 1 denotes with the reason of the information of feature entirely to predict *Y*, and gain ratio (GR) = 0 means with the purpose around nil relation among *Y* along with fr|Y. The GR works well features by smaller amount values whereas the Information Gain (IG)

$$IG(fr,y) = H(fr) - H(fr|Y) \tag{3}$$

The measure of uncertainty related to an indiscriminate variable is entropy (*H*). *H* (Fr) and *H* (fr/*Y*) are the entropy of Fr and, correspondingly, the entropy of successive observation, *Y*:

$$H(fr) = -\sum_{i} P(\mathrm{fr_i})\log_2\left(P(\mathrm{fr_i})\right) \tag{4}$$

The highest data benefit value is 1. A function with a high data gain is important. For each trait, IG is independently calculated, and the top *k* values of characteristics choose the required characteristics. This FS algorithm, based on a philtre, does not delete redundant features:

$$H(fr|Y) = -\sum_{j} P(y_j)\sum_{i} P(fr_i|y_j)\log_2\left(P(fr_i|y_j)\right) \tag{5}$$

3.2.3 Gini Index (GI)

[13] is a multivariate FS filter measurement algorithm aimed at determining the ability of a function to distinguish between groups. GI of each function can be calculated by defined *C* groups, as GI can take the utmost value of 0.5 for a twofold classification. There are smaller GI values for the other related characteristics. Each function of GI is independently calculated, and the top *k* features with the smallest GI are selected. It also does not delete unwanted functions, including IG:

$$GI(fr) = 1 - \sum_{i=1}^{C}\left[P(i|fr)\right]^2 \tag{6}$$

3.3 Filter Algorithm

The filter algorithm selects the rank features that are most elevated, and then the subset features selected can be used for some other classification.

3.3.1 Bat Algorithm (IBA)

Bat algorithm is known as the sound of echolocation produced by bats. Echolocation is characteristic sonar used by bats to locate predators and eliminate potential obstacles. Bats can produce louder sounds and produce echo with the intention of jumping back from the obstacles nearby. Therefore, a bat will measure how far from a function they are. Bats can find variation between an obstacle and a victim among the variety, even in total darkness. In order to the collection of features from the dataset, the bat's algorithm follows some basic rules [14–16]:

Bats flutter haphazardly by means of velocity denoted by vi at locations 'xi' by '$f_{\min}$' frequency, which is changeable loudness and wavelength AO to evaluate the optimal characteristics, which also facilitates an update of the pulse emission rates rϵ [0, 1], with respect to the nearness to the categorization accuracy, although its soundness is capable of diverging different ways. Let us focus on the varied usage of soundness from AO to lesser unchangeable value 'Amin'.

3.3.2 Initialization of Bat Population

The preliminary population features are randomly produced since original CKD dataset samples with information size *d* and *n* no. of bats, by consideration of lower and upper boundaries shown in:

$$x_{ij} = x_{\min,j} + rand(0,1)\left(x_{\max,j} - x_{\min,j}\right) \tag{7}$$

where $i = 1, 2, n$, $j = 1, 2, .d$, $x_{\min}$, j and $x_{\max}$,are lesser and higher limitations for feature j correspondingly.

3.3.3 Update Procedural Frequencies, Velocity and Remedy

The factor of frequencies can control the size of steps of feature selection remedy for BA. The frequencies in the lower and upper boundary are denoted by ($f_{\min}$ and $f_{\max}$). The speed of feature selection feature selection result is comparative towards frequency and novel result depending on its current velocity:

$$f_i = f_{\min} + \left(f_{\max} - f_{\min}\right)\beta \tag{8}$$

$$v_i^t = v_i^{t-1} + \left(x_i^t - x^*\right)f_i \tag{9}$$

$$x_i^t = x_i^{t-1} + v_i^t \tag{10}$$

where βϵ [0, 1] specifies arbitrarily creating a value, x* denotes the existing overall best feature selection solution. For neighbourhood search part of utilization, one feature selection result is chosen between the certain best solutions and the unsystematic walk is useful:

$$x_{\text{new}} = x_{\text{old}} + \varepsilon \overline{A^t} \tag{11}$$

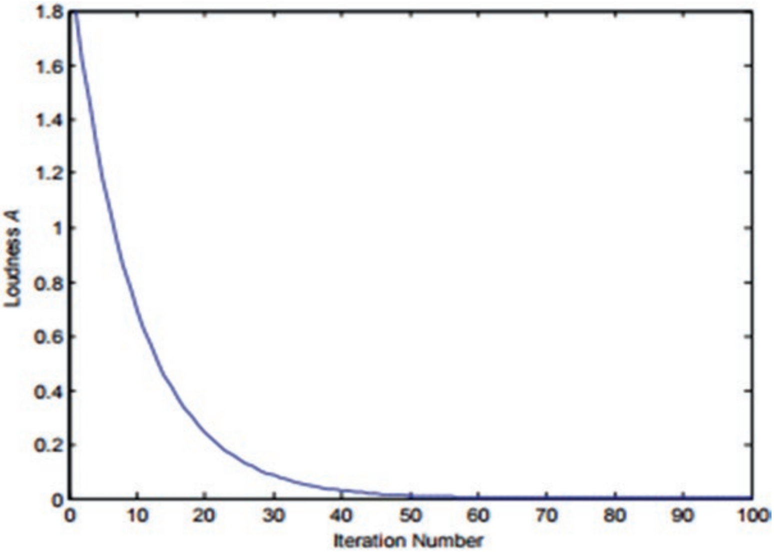

Fig. 3 Loudness *A*

where *t* [0, 1] is∈ε*A* is average loudness of each and every bat, arbitrary number and denotes way and strength of unsystematic walk.

3.3.4 Changing Rate of Pulse Emission and Loudness

Loudness *A* with pulse emission rate *r* is reorganized as (Fig. 3)

$$A_i^{t+1} = \alpha A_i^t \tag{12}$$

$$r_i^{t+1} = r_i^0 \left[1 - e^{-\gamma t}\right] \tag{13}$$

Where α and γ are constants. r_i^0 & Ai are features those includes of random values and A_i^0 be able to characteristically be [1, 2], while r_i^0 be able to characteristically be [0, 1].

Algorithm 1. Pseudo Code of BA

```
1. Intention of function: $f(x)$, $x = (x1,\ldots xd)t$
2. Initiate bat population $xi$ and velocity $v_{ii} = 1, 2, n$
3. Describe frequency of pulse $f_i$ at $x_i$
4. Initiate pulse rate ri with loudness $A_i$
5. when ($t <$ greatest iterations of frequency)
6. Produce latest solutions by fiddling with frequency, and by bringing up to date
   velocities and location/solutions.
7. F (rand > $r_i$)
```

```
8. Choose a feature selection solution between the best-selected features solution
9. Create a local feature selection solution with the best-picked features
10. end if
11. if (rand< $A_i$ and $f(xi) < f(x^*)$)
12. Recognize recent feature selection solutions
13. Raise ri,reduce $A_i$
14. end if
15. Find the position of ranking the bats and uncover updated recent best-selected
    features $x^*$
16. end while
17. Exhibit final feature selection outcomes.
```

3.3.5 Improved Bat Algorithm (IBA)

BA reavailable a critical algorithm that is capable of straightforwardly getting falls in local smallest on the majority of the functions. To solve this two major changes are applied to increase investigation and development ability of BA.

3.3.6 Inertia Weigh Factor Modification

In general, pulse emission rate r is used to control research and development of BA, and this element improves as it does until it reaches iteration (Fig. 4). At BA, steps 8 and 9 bear the localized BA hunt. The algorithm slowly decreases the production potential as iteration continues when step 7 is analysed. BA's research and development was affected by the weight of inertia factor. Then when the weight of inertia is higher, it is more like the quest overall. The effect of earlier velocity decreases gradually with the linear decreasing inertia weight factor. Consequently, BA's rate of growth is rising slowly as iterations continue:

$$w_{\text{iter}} = \frac{\text{iter}_{\max} - \text{iter}}{\text{iter}_{\max}} \left(w_{\max} - w_{\min} \right) + w_{\min} \tag{14}$$

where iter is the available iteration count, $\text{iter}_{\max}$ is maximum noumber of iterations and $w_{\max}$ and $w_{\min}$ are maximum and minimum inertia weight factors, respectively.

3.3.7 Adaptive Frequency Modification

In IBA, each feature of a solution is allocated a frequency from $f_{\min}$ to $f_{\max}$ individually:

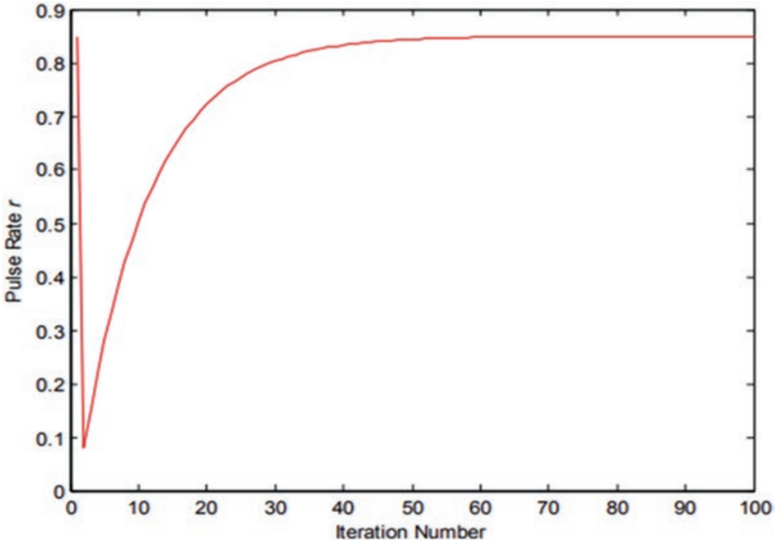

Fig. 4 Pulse emission rate *r*

$$diff_j = \sqrt{\left(x_{ij} - x_j^*\right)^2} \tag{15}$$

$$\text{Range} = \max\left(\text{diff}\right) - \min\left(\text{diff}\right) \tag{16}$$

$$f_j = f_{\min} + \frac{\sqrt{\left(\max\left(\text{diff}\right) - \text{diff}\left(j\right)\right)^2}}{\text{range}} * \left(f_{\max} - f_{\min}\right) \tag{17}$$

First, distances between solution *i* and the most excellent global feature selection solution are measured for each and every dimension, then neighbouring and furthest distance dimensions of selected features are allocated correspondingly to $f_{\min}$ and $f_{\max}$, and finally the frequencies of other feature dimensions differ in the $f_{\min}$ and $f_{\max}$ series with respect to their distances (Eq. (15)–approx. (17)). When we are analysing the figure (Fig. 5), the object of the solution's closest dimension to the largest is 1, and the farthest away is 3. First feature component of result increases gradually, at the same time as the third dimension of selected feature solution moves quickly. Else it moves from $f_{\min}$ to $f_{\max}$. Consequently, the dimensions are closer to the global feature selection solution. Velocity formulation (Eq. (9)) should be changed as follows:

$$v_{ij}^t = v_{ij}^{t-1} + \left(x_{ij}^t - x_j^*\right) f_j \tag{18}$$

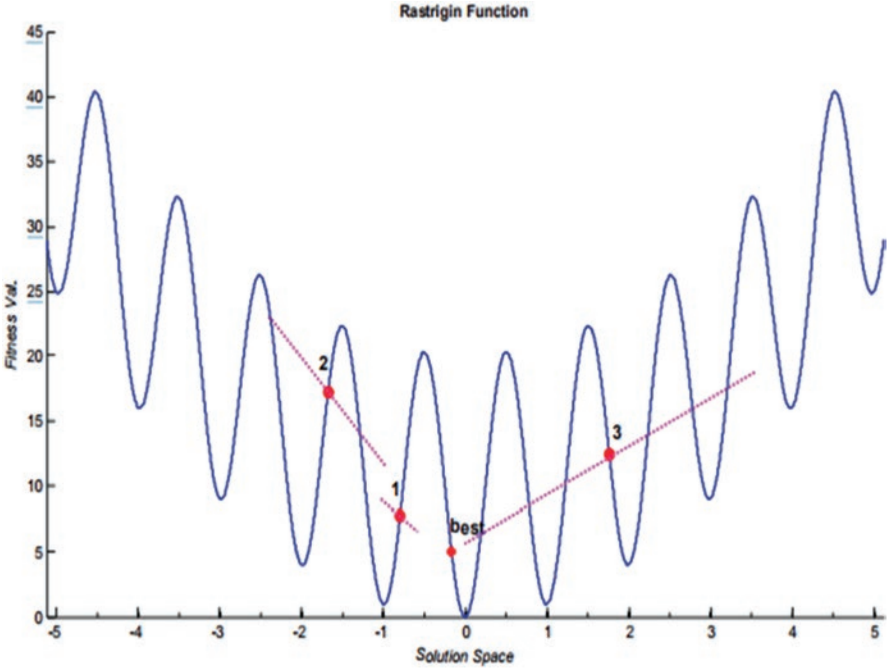

Fig. 5 Distribution of frequency

3.4 Embedded Algorithm

The embedded algorithm is underpinned with respect to SVM-t statistics to select the most preferred analytical feature in the CKD dataset. So, with the difference of trials, we are able to classify the main dissimilarity for specific genes between the nearest information points:

$$\left|t_j\right| = \left|\frac{\left(u_j^+ - u_j^-\right)}{\sqrt{\left(\left(s_j^+\right)^2 / n^+\right) + \left(\left(s_j^-\right)^2 / n^-\right)}}\right| \tag{19}$$

whereby $n+$ resp, $n-$) reavailables the number of supportive vectors that are formulated for class +1 (resp, −1). Compute mean u_j^+ (resp., u_j) and standard deviation s_j^+ (resp., s_j) based on typically supportive vectors of characteristic j-denoted category +1 (resp, −1) to evaluate the overall score of each feature.

4 Experimental Results

The classifier results were experimented via the use of the MATLAB tool with Intel Core 2 Duo Processor E7400 CPU (2.8 GHz Dual-Core, 1066 MHz FSB, 3 MB L2 cache) and 2 GB RAM.

4.1 Dataset Information

Glomerular filtration rate (GFR) can be calculated using creatinine serum, age, sex, size of the body, ethnic origin, etc. When a kidney's work is regular and polluted to a limit, unnecessary stuff will bring the blood elements up to elevated levels. Table 1 shows the stages of CKDs.

Table 2 symbolizes the CKD collection of information collected by UCI carrying 24 attributes along with another additional attribute for class (binary).

The confusion matrix is applied for narrating the classification performance algorithms through evaluating the performance metrics (Table 3).

The following metrics, such as classification accuracy (CA), accuracy, rate of errors, F-measure, accuracy and sensitivity, were utilized for the evaluation of the following: computed true positive (TP), false positive (FP), false negative (FN) and true negative (TN) and decision factor for the additional results:

True positive (TP) accurately refers to true information listed as true outcomes. True negative (TN) refers accurately to false information identified as false outputs. False positive (FP) refers to false information identified as real outputs, refers to real information, and is known as false outputs. False negative (FN) (accuracy of classification) refers to the algorithm used in the classification process in the CKD groups for diagnosis:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \tag{20}$$

Sensitivity specifies the correctness of evaluation of the target class's rate (Eq. 16):

Table 1 The stages of CKDs

Stages	Clinical features	GFR(mL/min/1.7 m^2)
I	Damage with normal or increased GFR	≥90
II	Damage with a mild decrease in GFR4	60–89
III	Moderate decrease in GFR	30–59
IV	Severe decrease in GFR	15–29
V	Kidney failure	<15 or dialysis

Table 2 The attributes of CKD of UCI

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

Table 3 Confusion matrix

Confusion Matrix		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN} \times 100 \tag{21}$$

Specificity relays in the direction of the test's ability in the approved manner detecting patients without a stipulated condition:

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100 \tag{22}$$

Precision also, namely, positive predictive value, is the fraction of same information between the recover information:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100 \tag{23}$$

F-measure is the average of precision and recall which is described as

$$\text{F} - \text{measure} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \times 100 \tag{24}$$

Figure 6 depicts the sensitivity and specificity results to classifiers like the NB, ANN and SVM. These results are measured in terms of the feature selected from the HFWE-FS algorithm; it concludes that these three classifiers perform better under HFWE-FS algorithm in Table 4. Nevertheless, the proposed SVM-HFWE-FS algorithm gives improved sensitivity results of 95.45% which is 13.08% and 14.77% high compared to ANN-HFWE-FS algorithm with SVM-HFWE-FS algorithm. Proposed SVM-HFWE-FS algorithm gives improved specificity results of 87.5% which is more efficient than the previous classifiers.

Figure 7 shows the precision and recall metrics evaluation to classifiers like the NB, ANN and SVM. The results are measured using the features selected from HFWE-FS algorithm; it proposed HFWE-FS algorithm with three classifiers that produce higher results than traditional classifiers. Proposed SVM-HFWE-FS algorithm gives improved precision results of 95.45% which is 5.54% and 6.66% more compared to ANN-HFWE-FS with SVM-HFWE-FS algorithm. Similarly proposed SVM- HFWE-FS algorithm gives improved F-measure outcomes of 95.45% which is 4.71% and 5.83% higher compared to ANN-HFWE-FS and NB-HFWE-FS methods correspondingly.

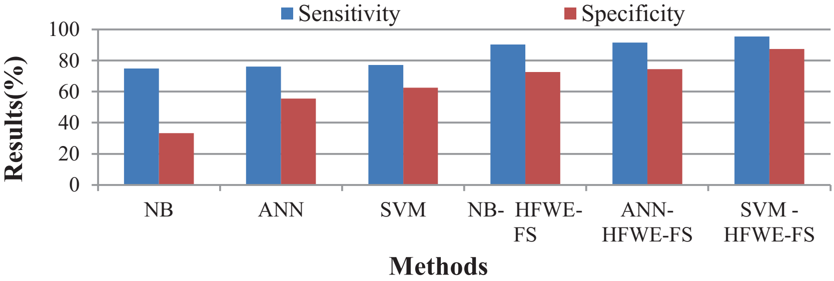

Fig. 6 Classifiers vs. sensitivity and specificity metrics

Table 4 Performance metrics vs. classifiers

	Results (%)					
Methods	Sensitivity	Specificity	Precision	F-measure	Accuracy	Error rate
NB-HFWE-FS	90.48	72.73	88.79	89.62	85.23	14.77
ANN-HFWE-FS	91.59	74.42	89.91	90.74	86.67	13.33
SVM-HFWE-FS	95.45	87.50	95.45	95.45	93.33	6.67
NB	75.00	33.33	81.82	78.26	66.67	33.33
ANN	76.19	55.56	80.00	78.05	70.00	30.00
SVM	77.27	62.50	85.00	80.95	73.33	26.67

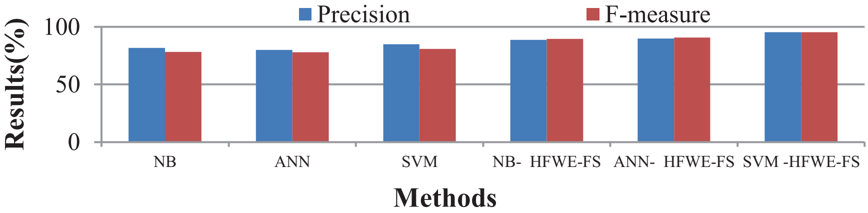

Fig. 7 Classifiers vs. precision and F-measure metrics

Figure 8 shows the accurateness and error rate metrics evaluation along with classifiers like the NB, ANN and SVM. The results are measured using the features selected from HFWE-FS algorithm; it proposed HFWE-FS algorithm with three classifiers that produce higher results than traditional classifiers. On the other hand, the anticipated SVM-HFWE-FS algorithm gives improved accuracy results of 93.33% which is 6.66% and 8.1% more compared to ANN-HFWE-FS and SVM-HFWE-FS algorithm. Similarly, proposed SVM-HFWE-FS algorithm gives reduced error rate results of 6.67% which is 6.66% and 8.1% less significant compared to ANN-HFWE-FS and NB-HFWE-FS classifiers correspondingly.

5 Conclusion and Future Work

A novel Hybrid Filter Wrapper Embedded (HFWE) is embedded in our paper along with Feature Selection (FS) algorithm to pick from the datasets the most favoured subset of features to predict CKD datasets.

Embedded, filter, and wrapper alongside the FS algorithm are utilized to minimize the feature attribute, and therefore SVM is also utilized in the classification of the attributes. The filter algorithm with four functions: Gini index (GI), gain ratio (GR), One R and Relief. The wrapper algorithm is implemented, and that is based on improved bat algorithms (IBA) to select the analytical feature from the wide-range CKD set of information. Embedded algorithm is executed with respect to the SVM-t statistics to choose the analytical attribute out of CKD dataset. SVM

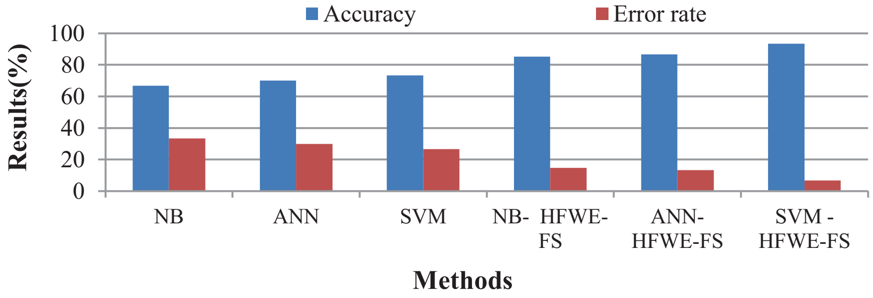

Fig. 8 Classifiers vs. accuracy and error rate metrics

classifier has been chased for the validity with reduced feature set. On a CKD dataset with 400 patients collected, an evaluation was carried out through the UCI machine learning repository. The main aim of this research is to evaluate whether CKD or non-CKD can be projected with even-handed precisions based on the chosen attributes from the HWF-FS algorithms. Performance was calculated in terms of six essential evaluation factors for categorization. On the other hand, the projected SVM-HFWE-FS algorithm gives improved accuracy results of 93.33% which is 6.66% and 8.1% more compared to ANN-HFWE-FS and SVM-HFWE-FS algorithm. Similarly, proposed SVM-HFWE-FS algorithm gives reduced inaccurate results of 6.67% that is 6.66% and 8.1% less significant in contrast to ANN-HFWE-FS and NB-HFWE-FS classifiers correspondingly. The same as explained from the outcomes, focus is more on the decreased features that are meant for recognizing CKD and thereby decreasing uncertainty, time saver with cost-effectivity. These set of information contain some unwanted and omitted values, so innovative classifier is needed to handle this issue efficiently. So it is left as future work.

References

1. Anderson, J., Glynn, L.G.: Definition of chronic kidney disease and measurement of kidney function in original research papers: a review of the literature. Nephrol. Dial. Transplant. **26**(9), 2793–2798 (2011)
2. Chen, Z., Zhang, X., Zhang, Z.: Clinical risk assessment of patients with chronic kidney disease by using clinical data and multivariate models. Int. Urol. Nephrol. **48**(12), 2069–2075 (2016)
3. Cho, B.H., Yu, H., Kim, K.W., Kim, T.H., Kim, I.Y., Kim, S.I.: Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artif. Intell. Med. **42**(1), 37–53 (2008)
4. Di Noia, T., Ostuni, V.C., Pesce, F., Binetti, G., Naso, D., Schena, F.P., Di Sciascio, E.: An end stage kidney disease predictor based on an artificial neural networks ensemble. Expert Syst. Appl. **40**(11), 4438–4445 (2013)
5. Go, A.S., Chertow, G.M., Fan, D., McCulloch, C.E., Hsu, C.Y.: Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. N. Engl. J. Med. **351**(13), 1296–1305 (2004)

6. Holte, R.C.: Very simple classification rules perform well on most commonly used datasets. Mach. Learn. **11**(1), 63–90 (1993)
7. Huang, M.J., Chen, M.Y., Lee, S.C.: Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert Syst. Appl. **32**, 856–867 (2007)
8. Karegowda, A.G., Jayaram, M.A., Manjunath, A.S.: Feature subset selection problem using wrapper approach in supervised learning. Int. J. Comput. Appl. **1**(7), 13–17 (2010)
9. Kathuria, P., Wedro, B.: Chronic kidney disease quick overview. IOP Publishing emedicine health, http://www.Emedicinehealth.com/chronic_kidney_disease/page2_em.htm#chronic_kidney_disease_quick_overview, 2016
10. Komarasamy, G., Wahi, A.: An optimized K-means clustering technique using bat algorithm. Eur. J. Sci. Res. **84**(2), 263–273 (2012)
11. Kumar, M.: Prediction of chronic kidney disease using random Forest machine learning algorithm. Int. J. Comput. Sci. Mob. Comput. **5**(2), 24–33 (2016)
12. Ladha, L., Deepa, T.: Feature selection methods and algorithms. Int. J. Comput. Sci. Eng. **3**(5), 1787–1797 (2011)
13. Norouzi, J., Yadollahpour, A., Mirbagheri, S.A., Mazdeh, M.M., Hosseini, S.A.: Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system. Comput. Math. Methods Med, 1–9 (2016)
14. O'Seaghdha, C.M., Lyass, A., Massaro, J.M., Meigs, J.B., Coresh, J., D'Agostino, R.B., Astor, B.C., Fox, C.S.: Risks score for chronic kidney disease in the general population. Am. J. Med. **125**(3), 270–277 (2012)
15. Salekin, A., dStankovic, J.: Detection of chronic kidney disease and selecting important predictive attributes, IEEE International Conference on Healthcare Informatics (ICHI), pp. 262–270, 2016
16. Priscila, S.S., Hemalatha, M.: Diagnosis of heart disease with particle bee-neural network. Biomed. Res. Spec. Issue, S40–S46 (2018)