UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Genetic Algorithm Ensemble Filter Methods on Kidney Disease Classification

Sharin Hazlin Huspi & Chong Ke Ting
School of Computing, Faculty of Engineering,
Universiti Teknologi Malaysia (UTM),
81310 Johor Bahru, Johor
Email: sharin@utm.my; keting.chong0610@gmail.com

Submitted: 6/9/2021. Revised edition: 7/10/2021. Accepted: 26/10/2021. Published online: 15/11/2021
DOI: https://doi.org/10.11113/ijic.v11n2.345

Abstract—**Kidney failure will give effect to the human body, and it can lead to a series of seriously illness and even causing death. Machine learning plays important role in disease classification with high accuracy and shorter processing time as compared to clinical lab test. There are 24 attributes in the Chronic Kidney Disease (CKD) clinical dataset, which is considered as too much of attributes. To improve the performance of the classification, filter feature selection methods used to reduce the dimensions of the feature and then the ensemble algorithm is used to identify the union features that selected from each filter feature selection. The filter feature selection that implemented in this research are Information Gain (IG), Chi-Squares, ReliefF and Fisher Score. Genetic Algorithm (GA) is used to select the best subset from the ensemble result of the filter feature selection. In this research, Random Forest (RF), XGBoost, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Naïve Bayes classification techniques were used to diagnose the CKD. The features subset that selected are different and specialised for each classifier. By implementing the proposed method irrelevant features through filter feature selection able to reduce the burden and computational cost for the genetic algorithm. Then, the genetic algorithm able to perform better and select the best subset that able to improve the performance of the classifier with less attributes. The proposed genetic algorithm union filter feature selections improve the performance of the classification algorithm. The accuracy of RF, XGBoost, KNN and SVM can achieve to 100% and NB can achieve to 99.17%. The proposed method successfully improves the performance of the classifier by using less features as compared to other previous work.**

Keywords—**Feature selection, genetic algorithm, random forest, k-nearest neighbor, XGBoost, support vector machine, naïve bayes**

I. INTRODUCTION

CKD is one of the top ten leading causes of death in the world [1]. In year 2015, the USA, Medicare expenditures for the CKD more than 64 billion States dollar [2]. However, [1] states that the signs and symptoms are vary, diagnosis is subjective and different from medical personnel. The epidemiological data are limited, the common lack of awareness and the poor access to the laboratory services probably cause the burden posed to kidney disease [2]. The number of nephrologists is relatively too small that not all the patients with CKD get the proper diagnosis. Therefore, technology like computer-aided diagnosis (CAD) is needed to assist the health care provider to diagnose CKD [1].

The machine learning methods were successfully applied to the biomedical field, especially to identify the presence of the disease and the risk appear according to the signs and symptoms [1]. The success of the application of machine learning mostly relies on the statistical approaches that can interpret a large set of the clinical datasets and give an extensive measure to the health condition of the patients. Some machine learning methods are popular to classify the disease which are KNN, NB, SVM and Artificial Neural Networks (ANN) [1], [3].

Feature selection plays a very important role in the medical field due to the high dimensional of the medical datasets [4]. For using all the high dimensional dataset needs more resources, money, time, and uncertainties [5]. Feature selection is used to select the subset of relevant features from the original high dimensional feature space [4]. A good feature selection can give performance improvement for the classifier by reducing the computing time and optimizing the use of the data in the dataset.

Thus, the goal of this research is to evaluate the proposed genetic algorithm ensemble filter feature selection (PGAEFFS) methods for selecting the informative and related features to classify CKD dataset. The research favours bringing the solution for the problem due to the features that are irrelevant and the unimportant which consist of fewer information in the

CKD dataset. Moreover, this research is contributing to information retrieval related to disease classification. It can help to increase the performance of the classification accuracy and help researchers to classify the disease using more effective dataset.

II. Literature Review

CKD is generally defined as disorders of the kidney affecting the structure and function in the heterogeneous form [5]. Currently, according to the international guidelines, a person is indicated as having the CKD if the function of the kidney decreased by showing the GFR is less than 60 mL/min per m^2, or the damage of kidney markers which identity with the presence of albuminuria, for at least three months [1]. The early detection able to facilitate the suitable diagnosis and treatment for the acute CKD is solicited for the augmented survivability [5]. Based on the Global Burden of Disease (2015), there are an estimated 1.2 million of people are dead due to kidney failure and it has increased significantly by 32% as compared with 2005. It is estimated that 5-10 million of people are dead due to kidney disease [2].

TABLE I. ABBREVIATION DESCRIPTION FOR THE FEATURES

Feature Abbreviation	Description	Permissible values
age	Age	Age in years
bp	Blood Pressure	in mm/Hg
sg	Specific Gravity	1.005, 1.010, 1.015, 1.020, 1.025
al	Albumin	0, 1, 2, 3, 4, 5
su	Sugar	0, 1, 2, 3, 4, 5
rbc	Red Blood Cells	normal, abnormal
pc	Pus Cell	normal, abnormal
pcc	Pus Cell Clumps	present, notpresent
bac	Bacteria	present, notpresent
bgr	Blood Glucose Random	in mgs/dl
bu	Blood Urea	in mgs/dl
sc	Serum Creatinine	in mgs/dl
sod	Sodium	in mEq/L
pot	Potassium	in mEq/L
hemo	Hemoglobin	in gms
pcv	Packed Cell Volume	in cells/cumm
wc	White Blood Cell Count	in cells/cumm
rc	Red Blood Cells Count	millions/cmm
htn	Hypertension	yes, no
dm	Diabetes Mellitus	yes, no
cad	Coronary Artery Disease	yes, no
appet	Appetite	good, poor
pe	Pedal Edema	yes, no
ane	Anemia	yes, no
ckd	CKD stage	ckd, notckd

The details for the feature abbreviation is described in Table I. According to [1], the 17 features selected which are *bp*, *sg*, *al*, *rbc*, *bgr*, *bu*, *sc*, *sod*, *pot*, *hemo*, *pcv*, *wc*, *htn*, *dm*, *appet*, *pe* and *ane* selected by using Correlation-Based Feature Selection (CFS) and AdaBoost had successfully improved the performance of KNN, NB and SVM for CKD classification by using a smaller number of features. Among all the three classifier, KNN classifier perform the best after applying feature selection with giving 0.981 accuracy rate, 0.980 recall rate and 0.980 f-measure rate [1]. Besides, the performance of the Random Forest, Regression Tree (CART) and SVM classifier for CKD classification had been improve after applying the union based of Info Gain, Gain Ratio, Chi Squared and Symmetric Uncertainty by selecting the 14 top ranked features which are *sg*, *al*, *rbc*, *pc*, *sc*, *hemo*, *pcv*, *htn*, *dm*, *bgr*, *bu*, *rc*, *appet* and *pe* [4]. By applying the wrapper feature selection which include recursive feature elimination, extra tree classifier and univariate selection had improved the performance of the XGBoost classifier for CKD classification by choosing *htn*, *hemo*, *al*, *dm*, *sg*, *pe*, *pcv*, *pcc*, *ane*, *appet* and *pc* as input. The XGBoost improved by giving the accuracy of 0.976, sensitivity of 1.0 and specificity of 0.917 [6].

Feature selection is mainly divided into filter, wrapper and embedded [7]. Filter feature selection selects the features based on their ranking and the features with the highest ranking will be selected without learning algorithm [4]. Filter method not only improves the performance of the classification method, but it also reduces the computer processing time. The disadvantage of this method is that it does not interact with the classifier algorithm. Besides, most of these methods are univariate, which means that the values of other attributes are do not into consideration. The filter can be divided into univariate filtration which include IG and Chi-Squares, and multivariate filtration which are ReliefF and Fisher Score [8].

Wrapper method applies the preselected induction algorithm for the feature selection process. The wrapper can give a better result than the filter method. However, the wrapper method has high computational cost as compared to the filter method and likely to overfit the data [8]. Sequential forward feature selection, sequential backward selection and genetic algorithm are the popular algorithms involved in wrapper feature selection [9].

III. Experimental design

In this section, the flow of research is briefly described to achieve the objectives. Overall, there are five important phases involved in this discussion.

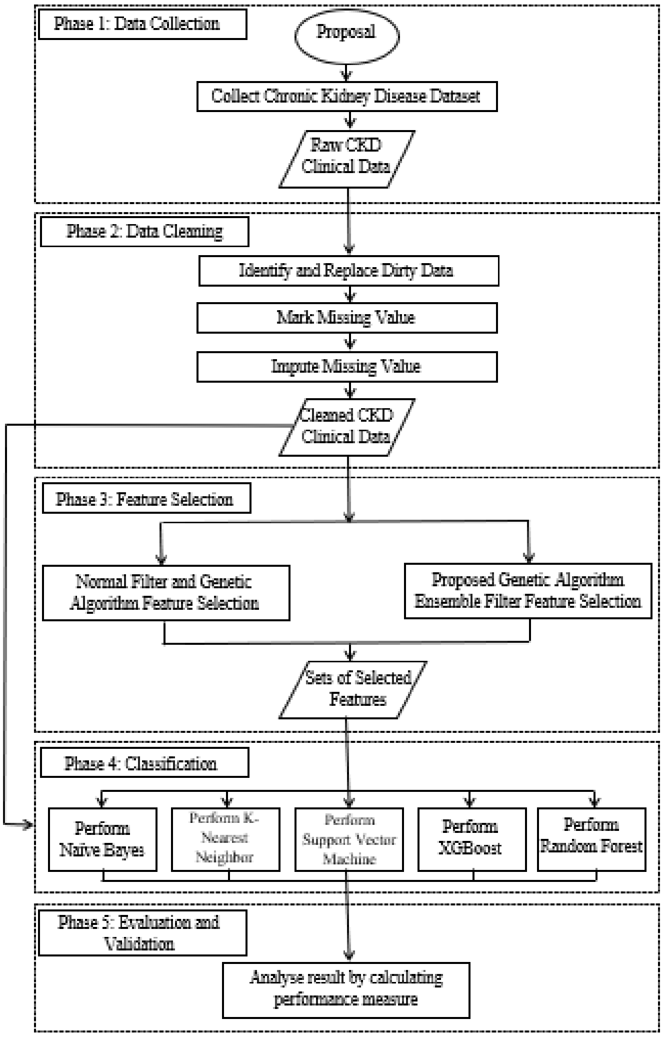
Phase 1: Data Collection
Proposal
Collect Chronic Kidney Disease Dataset
Raw CKD Clinical Data
Phase 2: Data Cleaning
Identify and Replace Dirty Data
Mark Missing Value
Impute Missing Value
Cleaned CKD Clinical Data
Phase 3: Feature Selection
Normal Filter and Genetic Algorithm Feature Selection
Proposed Genetic Algorithm Ensemble Filter Feature Selection
Sets of Selected Features
Phase 4: Classification
Perform Naïve Bayes
Perform K-Nearest Neighbor
Perform Support Vector Machine
Perform XGBoost
Perform Random Forest
Phase 5: Evaluation and Validation
Analyse result by calculating performance measure

Fig. 1. Experimental design of Research

Fig. 1 shows the overall working framework of the experiment for this research which including all the five phases involved.

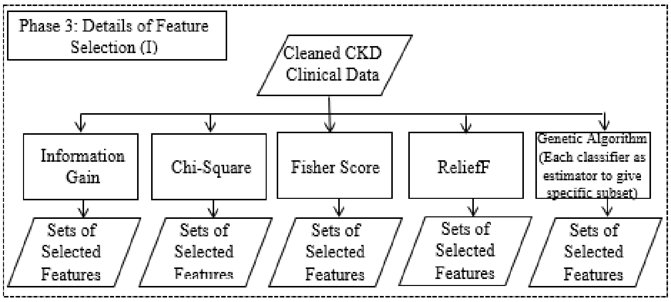
Phase 3: Details of Feature Selection (I)
Cleaned CKD Clinical Data
Information Gain
Chi-Square
Fisher Score
ReliefF
Genetic Algorithm (Each classifier as estimator to give specific subset)
Sets of Selected Features
Sets of Selected Features
Sets of Selected Features
Sets of Selected Features
Sets of Selected Features

Fig 2. Normal Filter and Genetic Algorithm Feature Selection

Fig. 2 demonstrate that the details of the normal filter and genetic algorithm feature selection which each of the feature selection are ran separately to generate different the features subset.

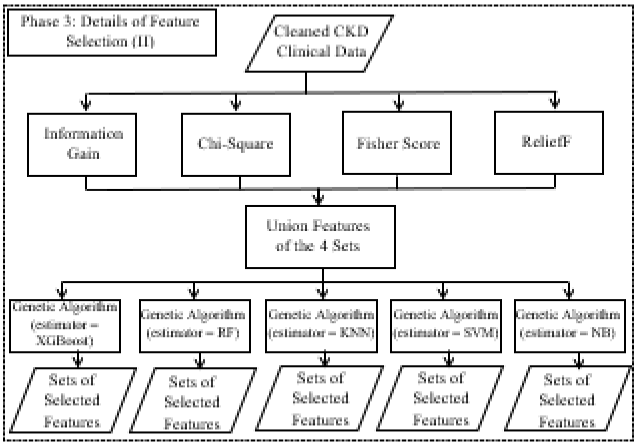
Phase 3: Details of Feature Selection (II)
Cleaned CKD Clinical Data
Information Gain
Chi-Square
Fisher Score
ReliefF
Union Features of the 4 Sets
Genetic Algorithm (estimator = XGBoost)
Genetic Algorithm (estimator = RF)
Genetic Algorithm (estimator = KNN)
Genetic Algorithm (estimator = SVM)
Genetic Algorithm (estimator = NB)
Sets of Selected Features
Sets of Selected Features
Sets of Selected Features
Sets of Selected Features
Sets of Selected Features

Fig. 3. Proposed Genetic Algorithm Ensemble Filter Feature Selection

Fig. 3 illustrate the details of PGAEFFS which the filter feature selection are employed to select the feature subsets. Then, the subset are ensembled to obtain the union subset to be fitted in into the genetic algorithm feature selection to obtain the best subset for each of the classifier.

A. Collection of Dataset

The chronic kidney disease (CKD) clinical dataset with 24 attributes and 400 records is downloaded and collected from the Kaggle in the .csv form. The attributes are the clinical records for each patient with a classification label that indicate whether the patients are suffer with CKD.

B. Data Cleaning

The dirty data in the raw dataset is first identified and replaced with the correct appropriate value by removing the '/t' invalid values in the record. Then, the missing value in the records is identified and labelled. The K-Nearest Neighbour imputation is used to fill in the missing values in the record. Besides, data normalization is carried out to make the range of values smaller.

C. Feature Selection

The pre-processed data is applied with the filter feature selection which are Chi-Square, Information Gain, ReliefF and Fisher Score to select the Top-6 features from each of the filter feature selection. Then the selected subsets are union to form a union of the features subset. Then, the union subset is used to undergo wrapper feature selection which is Genetic Algorithm feature selection to obtain the best features subset correspond to each of the classification algorithm.

Information Gain (IG) is used to measure the quality of the split by measuring the reduction in the entropy based on the given value of the random variable. The lower the entropy meaning that it is skewed probability distribution with unsurprising result and higher the purity of the samples [10]. IG rank the features by prioritizing based on the relevance of the features to determine different classes shows in Eq. 1.

$$H(X) = -\sum_i P(x_i) \log_2\big(P(x_i)\big), \tag{1}$$

let $P(x_i)$ to the prior probabilities value of X.

Eq. 2 shows the entropy of X after observing with another variable Y value.

$$H(X|Y) = -\sum_i P(y_i) \sum_i P(x_i|y_i) \log_2\big(P(y_i)\big) \tag{2}$$

where $P(y_i)$ represents the posterior probabilities of X given the values Y.

The amount of decrease of the entropy of X reflects additional information about X provided by Y is defined as IG. illustrate as:

$$IG(X|Y) = H(X) - H(X|Y) \tag{3}$$

According to the measurement, if $IG(X|Y) < IG\ (Z|Y)$, then it shows that feature Y is less correlated to feature X than to feature Z.

Chi-square test is used to measure the independence of the features towards the labels in the feature selection. Chi-square test is a statistical hypothesis categorised as chi-square distribution if the null hypothesis is true. [11]. The high score indicates that the high-dependent relationship of the features towards the labels. If the χ^2 value is smaller than the critical χ^2, it indicates that the correlation between the sample and the class is weaker and vice versa [12]. Eq. 4 is the chi-square calculation [12]:

$$\chi^2(r, c_i) = \frac{N[P(r,c_i)P(\underline{r},\underline{c_i}) - P\big(r,\underline{c_i}\big)P(\underline{r},c_i)]^2}{P(r)P(\underline{r})P(c_i)P(\underline{c_i})} \tag{4}$$

where N denotes the whole dataset and r shows the presence of the feature ($\underline{r}$ its absence) and c_i represents the class.

The Fisher Score measure the distance between the data points, the large inter-distance and small intra-distance is desired. The features that are more discriminant, it has higher Fisher score. Eq. 5 shows the calculation of Fisher Score.

$$SCF(F_i) = \frac{\sum_{j=1}^{C} n_{j\left(u_{i,j}-u_i\right)^2}}{\sum_{j=1}^{C} n_{j\,\sigma_{i,j}^2}} \tag{5}$$

where n_i μ_i $\mu_{i,j}$ μ_j $\sigma_{i,j}^2$ refer to the number of tuples that lay in the jth class, mean of the ith feature in the jth class, mean of the ith feature and the variance of the ith feature in the jth class, respectively. C refers to the total number of the class.

According to [13], ReliefF is the improve of the traditional Relief algorithm which used Euclidean distance during the measurement for instance distance and feature weighing that will fluctuat due to random acquired instances. ReliefF overcome the problem of ReliefF based on the Mean-Variance model to ensure the results obtained are accurate and stable. It depends on the number of neighbours which are set the parameter k to specify to compare with the k nearest hit and k nearest miss to overcome the noisy problem. The interpolation approach in which the class-conditional probability is set with the different function to overcome the missing value problem [13].

Genetic algorithm is a good heuristic method to use for the exploring of the feature space and able to generate many alternative subsets which include the most important features [19]. The genes of the chromosomes are represented as the number of features in the feature space. The genes "0" and "1" meaning the feature is selected and not selected respectively. Each chromosome in the GA is initialized randomly and the total number of the features is equal to the length of the chromosome [14]. The fitness function that used to optimize the performance of GA in this research is accuracy score. The fitness function score the performance of the feature subset and it is important for the subset selection. The cross-over is occurred to swap the section of the parents to form a new children chromosome that promote better result [15]. The mutation is occurred to replace specific number of features in the mutated subset by the features of the best subset from the previous generation. If the result obtained is lower than the given threshold, replace the number of features that have less weight with the highest important feature which does not exist in the subset of the best-found feature subset [15].

D. Classification

Five different machine learning algorithms are applied in this research to compare and observe that which algorithm perform better for the CKD classification. The machine learning algorithm that applied are Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbour and Naive Bayes. The best subset that are selected from the GA for each algorithm are fitted into their respective machine learning algorithm and undergoes the classification.

Random Forest are made up of classification tree, which is the simple model used for binary split on the variables to identify the prediction outcome. RF is made up of tree predictors thus each timber depends on the values of random vector sampled autonomously and having similar distribution of all the bushes in the forest. RF is first the tree bootstrap in figure out by taking out the distinctive samples of the data. Then each of the bootstrap samples gives an unpruned category tree. At each of the node, the predictors selection of the first-class split is not preferable, the arbitrary pattern attempt the predictors and from one of the variables, then select the acceptable breakup and new data is expected by adding up the predictions of n-trees and get the majority votes for types [15].

XGBoost is a tensile and new application of gradient boosting machines that have proved that it maximizes the computing power for the boosted tree algorithm. Boosting is the new model that implement ensemble technique that modify the errors that had been done by the existing model [16]. Gradient boosting is the algorithm that the model will be added recursively until there is no improvement observed. It is used to create the prediction of the residual of prior models, then add the prediction to decide the final prediction [16].

The basic idea of SVM is to seek to maximize the distance between two classes, the distance is known as the closest point (training point) [17]. The use of "kernel trick" in SVM, causing the distance between the particle and hyperplane able to be computed in nonlinear feature space that lacked the explicit transformation of the original descriptor [17]. The particle refers to the minimized true error by searching a hypothesis from the randomly selected test sample [17]. Furthermore, hyperplane refer to the training set which is the best in separating the positive data from the negative data in the dimensional space [17].

NB is a probabilistic classification algorithm that works according to the Bayes theorem with the features that assume it to be independent [19]. NB is a high scalability classifier but with the simple operation [1]. The classification problem can be solved effectively by using NB, although the variables do not have independence between each other in real life [1]. Assume that there is an instance of a vector of attributes with *n*-dimensional, then the Naïve Bayes will predict the vector to a class that has the highest posterior probability [1].

K-Nearest Neighbour is a non-parametric that is applied in statistical application [19]. KNN will store all the previous cases and classify the new cases given according to the similarity measure [20]. The basic idea of KNN is to calibrate a dataset, KNN will find a group of k samples that are nearest to the unknown samples [19]. The role of KNN is looking for the objects that are nearest to the exact point query or the majority vote [17].

The dataset is divided into 70:30 which are 70% of the dataset are used to train the model, while the rest of the 30% are used to test the performance of selected subset for each classifier.

E. Evaluation and Validation

The accuracy, precision, recall and f-measure are employed to observe and compare the performance for each of the machine learning algorithm.

Accuracy refers to the ability of classification algorithms to predict the classes of a dataset. Accuracy is used to identify how the classifiers have classified the dataset correctly [17]. The Eq. 6 illustrate the accuracy calculation.

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{6}$$

The recall is the sensitivity of retrieving the relevant instances. Eq. 7 illustrate the recall formula [17].

$$Recall = \frac{t_p}{t_p + t_n} \tag{7}$$

Furthermore, precision refers to a measurement of relevance that it retrieves information about relevant instances. The precision formula is illustrated in Eq. 8 [17]:

$$Precision = \frac{t_p}{t_p + f_p} \tag{8}$$

F-Measure is also called F-score. The function of F-score is to measure the accuracy of a test. It is the biased mean of recall and precision. The calculation of F-score is illustrated as Eq. 9 [26]:

$$F\ measure = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{9}$$

where t_p, t_n, f_p, f_n represent true positive, true negative, false positive and false negative respectively.

IV. RESULT AND DISCUSSION

The finding of this research discussed according to the accuracy, precision, recall and F-measure of the machine learning algorithm by using the original features sets and the feature subsets selected from chi-square, information gain, reliefF, fisher score, genetic algorithm, and the proposed genetic algorithm ensemble filter feature selections.

The accuracy of the proposed genetic algorithm ensemble filter feature selections, the traditional feature selections and the original dataset for Random Forest are shown in Table I.

TABLE II. CLASSIFICATION PERFORMANCE OF RANDOM FOREST WITH DIFFERENT FEATURE SELECTION METHODS

Feature Selection Method	Top 6			
	Accuracy (%)	Recall (%)	Precision (%)	F-measure (%)
Without Feature Selection	100	100	100	100
Chi-Square	99.17	98.67	100	99.33
Information Gain	100	100	100	100
ReliefF	94.17	90.67	100	95.10
Fisher Score	100	100	100	100
Genetic Algorithm	100	100	100	100
PGAEFFS	**100**	**100**	**100**	**100**

According to Table II, most of the selected features subset able to perform well with the 100% accuracy value, recall value, precision value and f-measure value in RF classifier. Among all the subset, the subset selected from the ReliefF perform the worst with only 94.17% accuracy, 90.67% recall and 95.10% F-measure. The RF able to perform well on most of the subset is due to the nature that the RF will select the most suitable features to build numbers of unpruned category trees, then obtain the priority vote from the output of decision tress as the final prediction result. Therefore, as long as the features that can give important information to differentiate the CKD and *nonckd* class is not eliminated, the RF is able to perform well by selecting the suitable features from all the features given to train the model.

TABLE III. CLASSIFICATION PERFORMANCE FOR XGBOOST WITH DIFFERENT FEATURE SELECTION METHODS

Feature Selection Method	Top 6			
	Accuracy (%)	Recall (%)	Precision (%)	F-measure (%)
Without Feature Selection	98.33	95.74	100	97.83
Chi-Square	96.67	94.67	100	97.26
Information Gain	100	100	100	100
ReliefF	94.17	90.67	100	95.10
Fisher Score	97.5	96	100	97.96
Genetic Algorithm	100	100	100	100
PGAEFFS	**100**	**100**	**100**	**100**

Based on Table III, the accuracy, recall, precision, and f-measure of XGBoost have been boosted to 100% by using the subset from the Information Gain, Genetic Algorithm and PGAEFFS feature selection as compared to the 98.33% when using the complete dataset. The is because the recall had been improved gradually from 95.74% to 100%. The selected features able to improve and boost the decision tree by constructing the best and clear path that can differentiate the *ckd* and *notckd* class during training. This is due to how the XGBoost works, that is by maximising the computing power of boosted tree algorithm which implement ensemble technique that modify the errors that had been done by the existing model. The exist of the irrelevant features in the subset will cause the confusion for the training of the XGBoost during the process of correcting the errors from the previous model, which causing the XGBoost unable to optimize it performance.

TABLE IV. CLASSIFICATION PERFORMANCE FOR K-NEAREST NEIGHBOR WITH DIFFERENT FEATURE SELECTION METHODS

Feature Selection Method	**Top 6**			
	Accuracy (%)	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Without Feature Selection	97.5	93.75	100	96.77
Chi-Square	99.17	98.67	100	99.33
Information Gain	100	100	100	100
ReliefF	94.17	90.67	100	95.10
Fisher Score	100	100	100	100
Genetic Algorithm	99.17	98.67	100	99.33
PGAEFFS	**100**	**100**	**100**	**100**

From the Table IV, the features subset selected from Chi-square, Information Gain, Fisher Score, Genetic Algorithm and PGAEFFS able to improve the performance of the KNN classifier from the original dataset without feature selection. This is because the original dataset consists of some of the features that are meaningless for the classification and causing the confusion for the training of the KNN models and unable to optimize the performance. The ReliefF perform the worst with only 94.17% accuracy, 98.67 recall and 95.10% f-measure. For KNN classification all the input features playing roles in the classification to increase the inter-distance and reduce the intra-distance between the centroids. If any of the irrelated features are selected, the inter-distance will be decreased and intra-distance increased, the model is not well trained, it will consequently reduce the performance of the classification methods.

TABLE V. CLASSIFICATION PERFORMANCE OF SUPPORT VECTOR MACHINE WITH DIFFERENT FEATURE SELECTION METHODS

Feature Selection Method	**Top 6**			
	Accuracy (%)	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Without Feature Selection	97.5	93.75	100	96.77
Chi-Square	99.17	98.67	100	99.33
Information Gain	100	100	100	100
ReliefF	94.17	90.67	100	95.10
Fisher Score	99.17	98.67	99.33	99.33
Genetic Algorithm	99.17	98.67	100	99.33
PGAEFFS	**100**	**100**	**100**	**100**

Table V shows that the selected features from the Chi-Square, Information Gain, Fisher Score, Genetic Algorithm and PGAEFFS able to improve the performance of SVM classifier from the original dataset. The 24 features from the whole complete dataset that contain some irrelevant attributes unable to provide a clear cut off point the differentiate the class during the training, the hyperplane constructed using this features subset do not have large margin that can easily differentiate the *ckd* and *notckd*. The selected features from the Information Gain and PGSEFFS perform the best with the 100% accuracy, recall, precision and F-measure. The hyperplane constructed using the selected features from Information Gain and PGAEEFS during training using these two subsets gives clear cut off point that have large margin between the two classes, therefore the classes can be predicted accurately when tested with the test dataset. While the features selected by ReliefF reduce the performance of the ReliefF classifier.

TABLE VI. CLASSIFICATION PERFORMANCE OF NAIVE BAYES WITH DIFFERENT FEATURE SELECTION METHODS

Feature Selection Method	**Top 6**			
	Accuracy (%)	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Without Feature Selection	95.83	90	100	94.74
Chi-Square	99.17	98.67	100	99.33
Information Gain	97.5	96	100	97.96
ReliefF	94.17	90.67	100	95.10
Fisher Score	80	68	100	80.95
Genetic Algorithm	95.83	93.33	100	96.55
PGAEFFS	**99.17**	**98.68**	**100**	**99.33**

Based on Table VI, the subset selected by PGAEFFS choosing the features subsets that perform the best for the NB classification. The subset selected from the Chi-Square and PGAEFFS giving the accuracy value 99.17%, recall value 98.68%, precision 100% and f-measure 99.33%. The selected subsets from Chi-Square and PGAEFFS maximize the different between the probability of classifying *ckd* class and *notckd* class in the training of NB classification. Therefore, the test data can be more clear and more accurate to classify into their respective class. However, the NB classification is not able to optimize to give 100% accuracy. The performance of NB classifier can be improved by increasing the recall value. The meaningless features in the dataset before feature selection causes the probability distribution between the *ckd* class and *notckd* class are close to each other during training. Therefore, when predicting with test the dataset the probability distribution calculated are quite near to the probability of the two classes and causing the confusion during the prediction. The selected subset from the ReliefF and Fisher Score reduce the performance of the NB classifier.

TABLE VII. SELECTED FEATURES AND CLASSIFIER PERFROMANCE FROM PGAEFFS AND EXISTING RESEARCH

Classifier	Research	Selected Features	Accuracy (%)
NB	Wibawa (2017) [1]	*bp, sg, al, rbc, bgr, bu, sc, sod, pot, hemo, pcv, wc, htn, dm, appet, pe* and *ane*	95.5
	PGAEFFS	**cad, su, htn, sg, pe, dm, hemo, al, pcv, rbc**	**99.17**
KNN	Wibawa (2017) [1]	*bp, sg, al, rbc, bgr, bu, sc, sod, pot, hemo, pcv, wc, htn, dm, appet, pe* and *ane*	95.8
	PGAEFFS	**su, htn, sg, pe, rc, dm, hemo, ba, al, pcv**	**100**
SVM	Wibawa (2017) [1]	*bp, sg, al, rbc, bgr, bu, sc, sod, pot, hemo, pcv, wc, htn, dm, appet, pe* and *ane*	96.3
	PGAEFFS	**pcc, cad, htn, sg, pe, rc, dm, hemo, ba, al, pcv**	**100**
RF	Shrivas (2018) [4]	*sg, al, rbc, pc, sc, hemo, pcv, htn, dm*	97.75
	Shrivas (2018) [4]	*sg, al, rbc, pc, sc, hemo, pcv, htn, dm, bgr, bu, rc, appet* and *pe*	98.50
	PGAEFFS	**pcc, cad, appet, htn, sg, pe, rc, sc, dm, ba, pcv**	**100**
XGBoost	Ogunleye (2018)	*htn, hemo, al, dm, sg, pe, pcv, pcc, ane, appet* and *pc*	97.60
	PGAEFFS	**cad, pc, sg, sc, dm, hemo, al, pcv**	**100**

The result showing that performance of all classifiers are improved after applying the proposed feature selection method. Besides, the proposed method in this research able to improve the performance of all the classifiers and it gives a better result as compared to the existing research result. Besides, the features subset that selected by the proposed method are having a smaller number of features than the existing research. Therefore, the proposed method is performed better for the CKD classification by improving the accuracy with using less features.

Based on all the existing research and experiments carry out, it can be concluded that every machine learning algorithm for *ckd* classification needs different subset of features to optimize the performance. Therefore, the filter feature selections that select the general features without considering the machine learning algorithm does not able to improve and optimize the performance of all the classification algorithms. Besides, the filter feature selections need manually identified the number of selected features, which is also one of the challenges to overcome and select the best subset for the classification.

Genetic algorithm that takes the machine learning algorithm performance in consideration during the process of feature selection can solve the issues that raised by filter feature selection as mentioned above. However, too large number of features with too many irrelated features may causing the genetic algorithm hard to identify the best subset and the less accurate subset will be obtained caused by the disturbance of the irrelated features.

Hence, the PGAEFFS that first remove some of the irrelevant features using the different filter feature selection methods, then union the selected features from each of the filter feature selections before the subset are fitted into the genetic algorithm feature selection. The proposed method able to resolve the issue that faced by the filter feature selections and genetic algorithm. The selected features from the proposed method able the give the best performance for all the five classification models. The RF, XGBoost, KNN and SVM can even be optimize by giving the 100% accuracy. For the NB even it does not optimize to give 100% accuracy, but it performs nearly perfect by giving 99.17% accuracy.

Among all the five classification methods, RF is the models that can perform the best for the CKD clinical dataset. It can perform well in most of the cases that with different selected features. This is because the RF will first select some of the important features form the input features to create a few numbers of decision trees with giving different path to classify the classes. Then, the prediction results from each of the decision tree are summing up to obtain the priority voting, then the most voted result from the decisions tree will be act as the result from the RF.

The NB is the classification algorithm that perform the poorest as compared to the other four classification algorithm. It is hard to optimize the performance of NB using this CKD clinical dataset. In most of the cases, the NB perform the poorest even though using different subsets. The NB algorithm unable to perform as well as other is because the classifying of NB is done by calculating the probability distribution of each category of features to the class and then calculate the probability of the *ckd* class and *notckd* class. The imbalance number of *ckd* and *notckd* class with the ratio of approximately 70:30 and some of the features that do not have clear cut off point to differentiate the *ckd* and *notckd* causing the probability of getting both classes are close to each other. This causing more difficult to increase the probability difference in classfying between *ckd* and *notckd*. Besides that, it might be due to the number of records is not large enough, since NB require large number of records to perform better. Therefore, we assumed that the algorithm is not well trained therefore the result is not as good as the other algorithm.

V. CONCLUSION

From this research, the PGAEFFS able to select the best features subset to improve the performance of the classification model. The RF, XGBoost, SVM and KNN are improved gradually and gives well classification performance, all these four classification models had been improved with giving the 100% accuracy, which mean that the models able to classify the CKD 100% accurately. It can be said that these four classification models are suitable to be employed to classify the CKD. The PGAEFFS had effectively removing the meaningless and redundant features to reduce the size of the dataset and subsequently improve the performance of the classification models. Among all the four classification models, XGBoost perform the best since it gives the 100% accuracy by using only 8 selected features which is only 33.33% of the original dataset. There is some improvement can be done to make this research become more meaningful. The improvement can be done is implementation of hyperparameter tuning for the genetic algorithm feature selection to optimize and stabilize the performance of the genetic algorithm. Besides, exploration with other alternative wrapper feature selection that can give better performance result and lower the computational cost. Lastly, the dataset can be implemented for multiclass classification to identify the stage of the CKD patient instead of just knowing whether there are having CKD since it can be categorized into five stages.

ACKNOWLEDGMENT

The authors would like to express their appreciation to the School of Computing, Universiti Teknologi Malaysia (UTM) for supporting this research.

REFERENCES

[1] Wibawa, M. S., Maysanjaya, I. M. D., & Putra, I. M. A. W. (2017, August). Boosted Classifier and Features Selection for Enhancing Chronic Kidney Disease Diagnose. *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, IEEE, 1-6.

[2] Luyckx, V. A., Tonelli, M., & Stanifer, J. W. (2018). The Global Burden of Kidney Disease and the Sustainable Development Goals. *Bulletin of the World Health Organization*, 96(6), 414.

[3] Kunwar, V., Chandel, K., Sabitha, A. S., & Bansal, A. (2016, January). Chronic Kidney Disease Analysis using Data Mining Classification Techniques. *2016 6th International Conference-Cloud System and Big Data Engineering*, IEEE, 300-305.

[4] Shrivas, A. K., Sahu, S. K., & Hota, H. S. (2018, April). Classification of Chronic Kidney Disease with Proposed Union Based Feature Selection Technique. *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, 26-27.

[5] Misir, R., Mitra, M., & Samanta, R. K. (2017). A Reduced Set of Features for Chronic Kidney Disease Prediction. *Journal Of Pathology Informatics.*

[6] M. Young. (1989). *The Technical Writer's Handbook*. Mill Valley, CA: University Science.

[7] Ogunleye, A., & Wang, Q. G. (2018, June). Enhanced XGBoost-based Automatic Diagnosis System for Chronic Kidney Disease. 2018 IEEE 14th International Conference on Control and Automation (ICCA), IEEE, 805-810.

[8] Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of Chronic Kidney Disease based on Support Vector Machine by Feature Selection Methods. *Journal of Medical Systems*, 41(4), 55.

[9] Villacampa, O. (2015). Feature Selection and Classification Methods for Decision Making: A Comparative Analysis. Doctoral Dissertation, Nova Southeastern University.

[10] El Aboudi, N., & Benhlima, L. (2016, September). Review on Wrapper Feature Selection Approaches. *2016 International Conference on Engineering & MIS (ICEMIS), IEEE*, 1-5.

[11] Brownlee, J. (2019, November 4). Information Gain and Mutual Information for Machine Learning. Retrieved from https://machinelearningmastery.com/information-gain-and-mutual-information/.

[12] Chitsaz, E., Taheri, M., Katebi, S. D., & Jahromi, M. Z. (2009). An Improved Fuzzy Feature Clustering and Selection based on Chi-Squared-Test. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 1, 18-20.

[13] Nissim, N., Moskovitch, R., Rokach, L., & Elovici, Y. (2012). Detecting Unknown Computer Worm Activity via Support Vector Machines and Active Learning. *Pattern Analysis and Applications*, 15(4), 459-475. Doi: 10.1007/s10044-012-0296-4.

[14] Urbanowicz, R. J., Meeker, M., Cava, W. L., Olson, R. S., & Moore, J. H. (2018). Relief-based Feature Selection: Introduction and Review. *Journal of Biomedical Informatics*, 85, 189–203. doi: 10.1016/j.jbi.2018.07.014.

[15] Ghareb, A. S., Bakar, A. A., & Hamdan, A. R. (2016). Hybrid Feature Selection based on Enhanced Genetic Algorithm for Text Categorization. *Expert Systems with Applications*, 49, 31-47.

[16] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. *Expert Systems with Applications*, 134, 93-101.

[17] Chen, T., & Guestrin, C. (2016, August). Xgboost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

[18] Zeynu, S., & Patil, S. (2018). Prediction of Chronic Kidney Disease using Data Mining Feature Selection and Ensemble Method. *International Journal of Data Mining in Genomics & Proteomics*, 9(1), 1-9.

[19] Patel, P., & Mistry, K. (2015). A Review: Text Classification on Social Media Data. *IOSR Journal of Computer Engineering*, 17(1), 80-84.

[20] Chandel, K., Kunwar, V., Sabitha, S., Choudhury, T., & Mukherjee, S. (2016). A Comparative Study on Thyroid Disease Detection using K-nearest Neighbor and Naive Bayes Classification Techniques. *CSI Transactions on ICT*, 4(2-4), 313-319.

[21] Thanh Noi, P., & Kappas, M. (2018). Comparison of Random Forest, K-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 imagery. *Sensors*, 18(1), 18.

[22] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2015). Classification of Heart Disease using k-nearest Neighbor and Genetic Algorithm. arXiv preprint arXiv:1508.02061.