IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 17, 2020, accepted November 21, 2020, date of publication November 26, 2020,
date of current version December 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3040650

Ensemble Feature Ranking for Cost-Based Non-Overlapping Groups: A Case Study of Chronic Kidney Disease Diagnosis in Developing Countries

SYED IMRAN ALI[1], HAFIZ SYED MUHAMMAD BILAL[1,2], MUSARRAT HUSSAIN[1], JAMIL HUSSAIN[1], FAHAD AHMED SATTI[1,2], MAQBOOL HUSSAIN[3], GWANG HOON PARK[1], (Senior Member, IEEE), TAECHOONG CHUNG[1], AND SUNGYOUNG LEE[1], (Member, IEEE)

[1]Department of Computer Science and Engineering, Kyung Hee University, Yongin 17104, South Korea
[2]School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[3]Department of Software, Sejong University, Seoul 05006, South Korea

Corresponding authors: Sungyoung Lee (sylee@oslab.khu.ac.kr) and Taechoong Chung (tcchung@@khu.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program supervised by the Institute for Information and Communications Technology Promotion (IITP) under Grant IITP-2017-0-01629, in part by the IITP Grant funded by the Korean Government (MSIT) under Grant 2017-0-00655, and in part by the Ministry of Science and ICT (MSIT), South Korea, through the Grand Information Technology Research Center Support Program under Grant IITP-2020-0-01489 supervised by the IITP under Grant NRF-2016K1A3A7A03951968 and Grant NRF-2019R1A2C2090504.

ABSTRACT Chronic kidney disease (CKD) is one of the leading medical ailments in developing countries. Due to the limited healthcare infrastructure and the lack of trained human resources, the CKD problem aggravates if it is not addressed in its earlier stages. In this regard, the role of machine learning-based automated diagnosis systems plays a vital role to deal with the CKD problem. In most of the studies conducted on the automated CKD decision modeling, the main emphasis is given to enhancing the predictive accuracy of the system. In this study, we focus on the applicability challenges of automated decision systems taking CKD diagnosis as a case study within the purview of developing countries. In this regard, we propose a cost-sensitive ensemble feature ranking method that takes a more realistic approach to group-based feature selection. Two candidate solutions are proposed for group-based feature selection to meet different objectives. Subsequently, both the candidate solutions are combined into a consolidated solution. It is pertinent to note that it is one of the first studies in which cost-sensitive ensemble feature ranking for non-overlapping groups is successfully demonstrated to achieve the stated objectives i.e. low-cost and high-accuracy solution. Based on an extensive set of experiments, we demonstrate that a cost-effective and accurate solution for the CKD problem can be obtained. The experimentation includes 7 well-known classification algorithms and 8 comparative feature selection methods to show the efficacy of the proposed approach. It is concluded that the applicability of the automated CKD systems can be enhanced by including the cost consideration into the objective space of the solution formulation. Therefore, a trade-off solution can be obtained that is cost-effective and yet accurate enough to serve as a CKD screening system.

INDEX TERMS Ensemble feature ranking, cost-based feature selection, threshold selection, filter methods.

I. INTRODUCTION

Chronic kidney disease (CKD) is a healthcare problem with serious consequences that is characterized by a gradual loss of kidney function over time. CKD is generally defined as abnormalities in the structure or function of the kidney or a decrease in Glomerular filtration rate (GFR) <60 ml/min/1.73 m^2 for 3 months [1]. The main function of the kidney is to filter out the excessive waste in the body along with balancing the body's fluids [2]. In the advanced

The associate editor coordinating the review of this manuscript and approving it for publication was Shiping Wen.

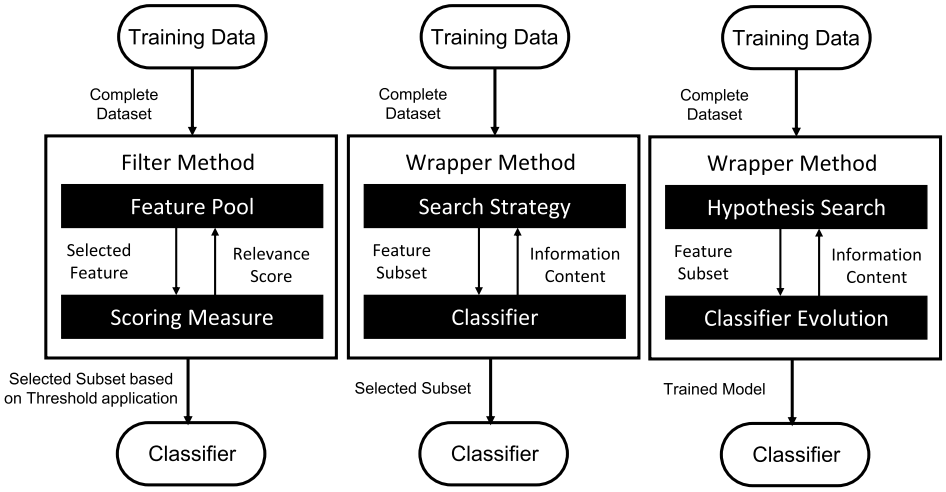

FIGURE 1. Feature selection approaches; (a) filter (b) wrapper, and (c) embedded.

TABLE 1. Nomenclature.

Notation	Description
O	Observed value in chi-squared statistic
E	Expected value in chi-squared statistics
W	Feature weight vector in Relief
TG	A particular test group
F	Consolidated feature score list
$\hat{f}$	Member feature in a particular test group TG
f	Member feature in the original feature set
$SU(A, B)$	Symmetric Uncertainty of two features
$ScoreTG(.)$	Relevance score of a particular TG
$TGcost(.)$	Accumulated cost of multiple test groups
$GCv(.)$	Cost of a particular TG
FeatureCost(.)	Accumulated cost of multiple features
$FCv(.)$	Cost of a particular feature
$Cost(.)$	Cost value of a particular feature
$RelativeRank(.)$	Localized rank of a feature w.r.t its TG

TABLE 2. Abbreviations.

Abbreviation	Full Form
CKD	Chronic Kidney Disease
GFR	Glomerular Filtration Rate
KFRS	Kernelized Fuzzy Rough Sets
SVM	Support Vector Machine
UCI	University of California at Irvine
ANN	Artificial Neural Networks
XGBoost	Extreme Gradient Boosting
LG	Linear Regression
RF	Random Forest
LASSO	Least Absolute Shrinkage and Selection Operator
GA	Genetic Algorithm
CART	Classification And Regression Trees
NB	Naive Bayes
KNN	K-Nearest Neighbor
PCA	Principal Component Analysis
TG	Test Group
DFS-CT	Direct Feature Selection - Combine Threshold
DFS-TC	Direct Feature Selection - Threshold Combine
PKR	Pakistani Rupees
SU	Symmetric Uncertainty
FSF	Feature Scoring Function

stages of kidney deterioration, bodily waste builds up that in turn impair the regulation of blood pressure, red blood cell creation, and the formation of bones, with life-threatening consequences. In case of severe kidney damage, the available options are in terms of renal replacement therapy or kidney transplant, where the latter is not a readily available treatment option, the former affects the overall quality of life while providing a temporary solution [3]. Therefore, it is of great value to diagnose CKD disease earlier in its formative stages.

Frequently used symbols and abbreviations are provided in Table 1 and Table 2.

CKD is one of the main causes of death in both developed and under-developed countries. It is estimated that around 1 million people succumbed to CKD related illnesses in 2013 [4]. Although CKD is a global scale concern, it affects the population in developing countries rather disproportionately [4]. It is well-documented that CKD is a highly prevalent disease in developing countries, one out of every ten persons are suffering from CKD related ailments in the South Asian region e.g., India, Bangladesh, Pakistan, Nepal, Bhutan, Sri Lanka, and Afghanistan. The CKD incidence and prevalence are attributed to a number of factors such as environmental, ethnic, socioeconomically, and rural-urban differences [5]. In a recent study on CKD in Pakistan, the regional patterns of CKD prevalence are contrasted with that of the developed countries and it is concluded that there is a similarity in the overall trends [5]. It is also reported that a large number of patients face sub-optimal outcomes in dealing with CKD due to severe economic hurdles [6]. Therefore, in developing countries, the population living under the poverty line is unfavorably situated to benefit from the advanced early-stage CKD screen techniques due to myriad factors such as cost of diagnosis, limited infrastructure in rural settings, and the lack of trained human resource, among others [7].

In this regard, a number of machine learning-based techniques are proposed to assist in CKD-related

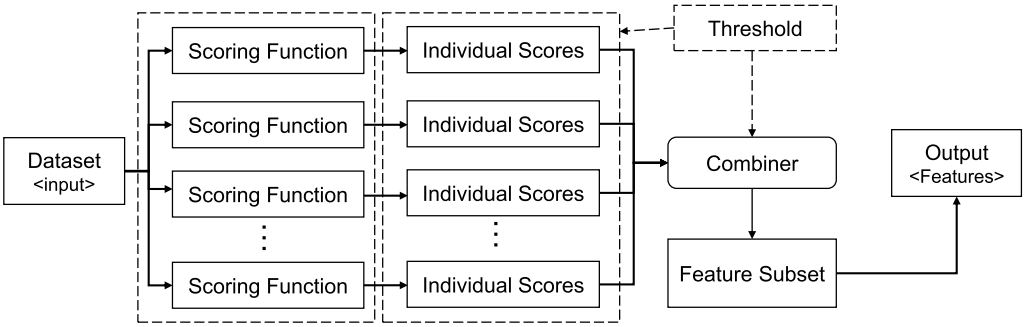

FIGURE 2. The schematic diagram for an ensemble-feature ranking approach [8].

diagnosis [9]–[13]. The overall objective of the CKD diagnosis system is to accurately and reliably diagnose patients through automated clinical decision support systems. These studies leverage a number of key indicators such as hypertension, diabetes mellitus, blood pressure, albumin, creatinine, among others, to flag a high-risk patient. The significance of such automated screening systems is to provide timely and cost-effective recommendations without overburdening the healthcare system. Most of the CKD diagnosis systems are designed to capture subtle patterns with higher accuracy that indicate the onset of the disease, thereby, improving the efficacy of the treatment in the early stages of the disease.

In the domain of healthcare, the interpretable machine learning modeling approaches are generally preferred as the user can inspect the underlying inferencing of the models [14]. It is noted by Itani *et al.* [15] that physicians are generally more appreciative of the decision support systems that take into account the operational considerations of the medical field. In this regard, predictive accuracy is one among many other considerations that affect the decision-making process. The cost of machine learning models is an important factor in operationalizing the decision support systems, especially in rural settings where infrastructure is scarce [16]. In this regard, it is of paramount concern that machine learning techniques that preserve the semantics of the data are used in modeling the decision process. Feature selection is a preprocessing technique that generally tends to increase the predictive accuracy of the machine learning models. These techniques are used in a variety of application domains such as cyber-security, business, biological data processing [17]. In this regard, the application of feature selection in the medical domain is two-folds i.e. it not only reduces the overall dimensionality of the problem but also identifies the salient factors in decision modeling. Furthermore, the speed of model construction is increased while at the same time the generalization of the model is also increased. Feature selection techniques are generally divided into three categories as shown in Figure 1, i.e. filter-methods, wrapper methods, and embedded methods. The filter methods generally tend to employ univariate statistical measures to evaluate the usefulness of the feature. Subsequently, features are ranked according to their relevance and, a subset of features is selected based on a user-defined threshold value. In the case of wrapper approaches a learning algorithm is used in the feature set optimization process where a subset of features is selected as a final solution. Embedded methods tend to select informative features in the process of model induction and hence embedded methods are implicitly used by some of the modeling techniques such as the C4.5 decision tree model.

Recently, ensemble-feature selection techniques have reported promising results as compared with non-ensemble techniques [18]–[20]. In ensemble techniques, multiple feature evaluation measures or the same measure on multiple data subsets can be used in parallel to provide a more comprehensive and robust evaluation. In the case of ensemble-methods, special consideration is given to the diversity and stability among the individual candidate solutions. Similarly, a consolidated feature ranking is obtained by combining multiple individual solutions, where each candidate solution corresponds to a specific evaluation technique used in constructing the ensemble i.e. an element of the ensemble. Two key considerations in ranking-based feature selection techniques are the feature weight-age technique and the threshold value selection. This study focuses on ensemble-based feature ranking along with threshold selection heuristic. A schematic diagram of the ensemble-feature ranking approach is shown in Figure 2, where two alternate scenarios are depicted i.e. applying threshold before combining individual candidate solutions and threshold application after combining results of the candidate solutions. Furthermore, the feature weight-age in the schematic illustration refers to a heterogeneous case.

A number of studies have reported promising results on the application of feature selection techniques to the CKD diagnosis problem. In these techniques both filter and wrapper based methods are used for selecting a set of salient features [9]–[13], [21]. In this case, it is demonstrated that a small set of highly predictive features can yield accurate and generalizable classification models. It is pertinent to mention that most of the studies in the domain of medical diagnosis assume that the cost of data acquisition is fixed

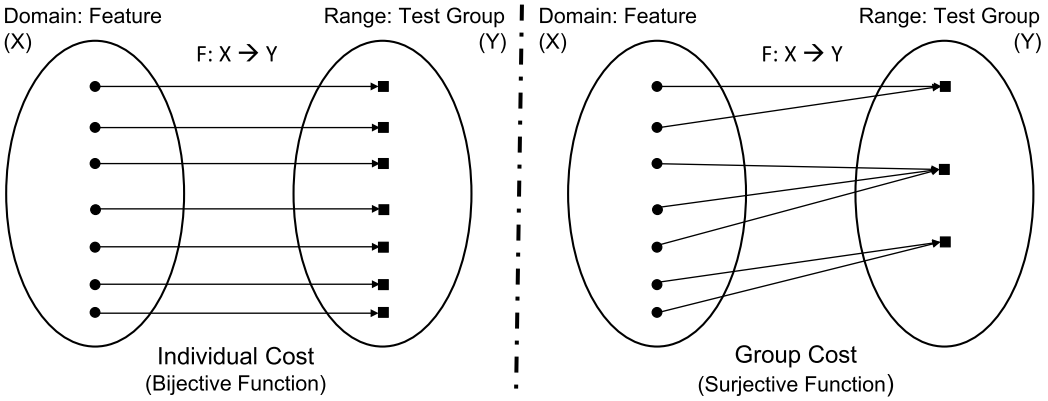

FIGURE 3. Cost-assignment based on the individual and group-based formulation.

across the feature space i.e. features' have the same cost, therefore, the cost is not taken into consideration in designing a diagnosis solution.

The notion of cost can be broadly defined as a contributing factor in the candidate solution that diminishes its desirability. In this regard, the cost can take a number of forms such as economic cost, the degree of risk in terms of side-affects associated with a particular medical procedure, the computational time required to process a sample as in the case of medical imaging, the availability of specialized medical equipment's, among others. Data acquisition cost is one of the operational aspects that has recently received attention from the machine learning community. In this regard, a few studies have reported the cost consideration in the CKD diagnosis problem [8], [22]. It is demonstrated that a cost-effective classification model aided by feature selection techniques can be generated for the CKD diagnosis problem having a reasonable predictive accuracy. In the aforementioned studies, the operating assumption is that the features are mutually exclusive in terms of cost assignment. Hence, the cost factor is associated with each feature on an individual basis. This assumption may not hold in a case where the data acquisition process is incumbent on medical tests where data for one or more features can be derived from a single medical test. A pictorial representation of individual and group tests is depicted in Figure 3. In the case of individual-cost formulation, it is inherently assumed that each test group would provide data for a single feature. Therefore, the number of test groups would be equal to that of features in the dataset. Whereas, in the case of group-cost formulation a small set of test groups are identified that would yield data for all the required features in the dataset. Therefore, the cost factor associated with a test group is shared among the features of the given test group. For example, a urine analysis test covers a wide range of features such as specific gravity, pus cells, red blood cells, and puss cells clumps. The aforementioned medical test varies in terms of the cost incurred due to a number of factors e.g., the type of the patient's medical insurance, government subsidy (in case of public hospitals or laboratories), and charges set by private laboratories. This study adopts the group-cost formulation, as it is more representative of the operational aspects of the CKD diagnosis. Furthermore, the application of cost is used in a consistent manner across the feature space in designing the solution i.e. the cost factor is based on medical tests conducted in public hospitals under an urban setting while the differences in data acquisition cost between public and private testing laboratories, and the laboratories in an urban and rural setting are not catered in this study.

The main motivation of this study stems from the need to take into account the operational and practical aspects of machine learning applications. In this regard, the selected problem is of high interest as it blights the developing countries over a long period. As machine learning-based screening and diagnosis applications provide a workable solution to the many problems faced by the healthcare sector, therefore, recently a lot of scholarly work is reported on developing disease diagnosis and management systems for the CKD [9], [10], [12], [13], [21], [22]. The current study is in the continuation of the scholarly work performed on the CKD diagnosis problem. The main objective of this study is to investigate and demonstrate the applicability of a holistic approach that takes into account both the predictive accuracy and operational cost of the machine learning solution. For a screening application, the overall objective is to reduce the cost of the application with a reasonable degree of reliability i.e. predictive accuracy.

This study deals with a number of research questions within the purview of the CKD diagnosis solution design:

1) How to assign an economic cost factor to a feature in a group formulation?
2) How to select a threshold value in a cost-effective manner?
3) How to apply a threshold to the ranked features?

The key contributions of this study are as follows:

- This study demonstrates an effective cost-sensitive feature ranking methodology for non-overlapping feature groups

- This study demonstrates an effective mechanism through which both cost and relevance score can be assigned to individual features and test groups, respectively
- This study provides an empirical evaluation of different threshold application and subset combination approaches within the scope of the CKD diagnosis problem
- It is one of the first studies on CKD diagnosis in which the economic cost factor is utilized from the perspective of a developing country

The rest of the paper is organized as follows: Section II deals with a detailed literature review within the purview of relevant studies on the CKD problem. The proposed methodology is discussed in Section III. Section IV deals with extensive experimentation of the proposed and comparative techniques. Finally, discussion and conclusions are provided in Section V and VI, respectively.

II. RELATED WORK

Feature selection is one of the key data preprocessing techniques generally used for enhancing the robustness of the machine learning classifiers [20], [23], [24]. Gu *et al.* [23] proposed a wrapper-based feature selection method. This approach used kernelized fuzzy rough sets (KFRS) for evaluating a set of candidate solutions generated through a memetic algorithm. In this study, it is demonstrated that irrelevant features induce a computational burden on the classification models. Therefore, resulting in sub-optimal models with reduced accuracy. The aforementioned evaluation function produced consistent scoring of the solutions while dealing with uncertainty and noise in the data. It is also demonstrated that the memetic search mechanism successfully obtained highly accurate solution and search success as compared with other meta-heuristic methods in a comparatively less time. Li *et al.* [24] proposed an ensemble of OS-extreme learning machine with binary Jaya (BinJaya)-based feature selection approach for the assessment of real-time transient stability of power systems. It is demonstrated that the proposed BinJaya approach is capable of selecting an optimal set of features from the entire feature space of the phasor measurement units data. This approach also used the KRFS technique as the class separability criterion. It is demonstrated that the BinJaya algorithm selected a set of 7 features that are highly predictive of the target concept from an original set of 33 features. The aforementioned studies show that feature selection techniques are generally used to enhance the predictive capabilities of the classification models.

As the real-world data are stored without having any specific application in mind, therefore, the data need to be adapted for the classification task e.g., removing irrelevant attributes. In this regard, one of the important aspects reported in literature is to identify useful features that not only enhance the predictive performance of the model but also provide the decision-maker with a set of important predictors for knowledge management [17]. In the case of the cost-sensitive feature selection framework, a feature is not only associated with value but also with a cost. Therefore, the main objective of this study is to extensively investigate the role of cost-sensitive feature selection within the scope of the CKD diagnosis problem as a case study. In this regard, this section deals with the broad categorization of feature selection techniques, some of the key studies on the decision modeling for the CKD problem, and some of the promising approaches in the direction of ensemble feature selection. Feature selection techniques are generally divided into three types i.e., filter, wrapper, and embedded methods. Filter methods are one of the widely used techniques for feature selection [25]. As a dataset may contain irrelevant features that may result in an over-fitted classification model, therefore, filter methods assign a relevance score to each feature in the dataset. This relevance score is generally based on statistical or information-theoretic measures such as information gain, gain ratio, fisher score, chi-square, t-test, inconsistency criterion, among others. In this regard, an ordered feature list is obtained based on univariate analysis. A subset of features can be drawn from this list by applying a threshold value. The automatic selection of a threshold value is a non-trivial task [26], [27], therefore, in most cases either the decision-maker is tasked with an appropriate threshold value selection or a fixed set of features are selected based on an apriori threshold selection e.g. top 10% features [28]. Filter methods are generally preferred for large datasets as they incur a less computational cost as no classifier is included in the feature scoring process. One of the key limitations of these methods is ignoring feature interaction. Methods that take into account pair-wise feature interaction, tend to take relatively more computational time as compared with the baseline filter methods. In this regard, wrapper approaches are suitable for capturing feature interaction. Wrapper methods are generally more accurate as they involve classifier in the solution evaluation stage, furthermore, feature interaction is addressed by generating multiple candidate feature subsets. The candidate subsets are iteratively evolved using powerful searching techniques such as evolutionary algorithms, randomized hill-climbing, beam search, branch-and-bound methods, simulated annealing, among others. The resultant feature subsets are more compact, account for feature interaction, and feature redundancy at the expense of computation time. Wrapper methods tend to produce less generalizable feature subset solutions as compared with filter methods, as the latter tends to rely on the intrinsic characteristics of the dataset while the former relies on the inductive bias of a particular classifier. The third category of feature selection involves embedded techniques in which the search mechanism is guided by the model creation process. In the course of building a model from training data, a nested subset evaluation is performed to select a suitable set of features for building the model [25]. A set of features are selected in the process of optimizing a classification model. Embedded techniques are computationally less expensive as compared with wrapper methods,

TABLE 3. Summarized comparison of individual feature selection approaches.

Type	Methodology	Merits	Demerits
Filter	All the features in a dataset are scored based on their relevance in predicting the class label. Generally, feature-feature interaction is not taken into account. A threshold value is required to select the final solution	• Generally less computationally expensive • Comparatively more generalizable • Suitable for very large datasets	• Generally, less accurate than the wrapper methods • Feature redundancy is difficult to address
Wrapper	Candidate feature subsets are generated and evaluated in an iterative manner. Over a number of predefine generations (or if a predefined accuracy objective is met), the candidate feature subset obtained thus far is selected as the final solution	• Generally, more accurate on small to medium size datasets • Feature-Feature interaction is captured • Generally, the resultant solution is more compact than the filter methods	• Computationally more expensive • Comparatively less scalable than the filter methods
Embedded	Feature selection is an internal operation within the classification model construction approach. A set of features are selected to form a more robust classification model	• Produces a highly efficient solution for a particular modeling approach • Generally less expensive than the wrapper methods	• Comparatively less generalizable as compared to both filter and wrapper methods • Generally more expensive than the filter methods

while the results produced are tightly coupled with a specific classification model and hence are less generalizable. Table 3 provides a summarized comparison of the main feature selection approaches.

Ensemble methods are frequently reported in the feature selection literature [18]–[20], [29], [30]. These ensemble techniques have taken inspiration from ensemble model learning where weak learners provide intermediate labels that are subsequently combined into a final recommendation. As the ensemble model learning has shown promising results on classification and regression problems, similar results can be observed on complex feature selection problems. In the case of the feature selection problem, generally, there are two approaches for building ensemble solutions i.e. ranking based approaches and subset based approaches. As feature selection methods can either produce an ordered ranked list or a final feature subset, a similar case is observed for ensemble feature selection approaches as well based on the nature of the base feature selection technique. For example, if the base methods in an ensemble tend to produce a ranked list of features, then the ensemble approach would produce a consolidated ordered ranked list. Generally, ensemble feature selection methods are divided into homogeneous and heterogeneous approaches. In the case of a homogeneous approach, the dataset is horizontally divided and the same feature selection method is applied to each data partition. While in the case of a heterogeneous approach, multiple feature selection methods are applied to the same data partition. In this study, we explore the heterogeneous ensemble approach that is relatively more common [31]. As our proposed ensemble approach deals with both feature ranking and feature subsets, therefore, we focus on these two aspects of the ensemble feature selection methodology. Table 4 provides a summarized comparison of the two aforementioned techniques for ensemble approaches based on the base feature selection methods. Subset-based methods produce results in a partial list of features i.e., features present in the partial list are considered elements of the final solution while the rest of the features are discarded for any further processing. In this case, each base subset method produces a partial list. This partial list may or may not be ordered, therefore, the relevance of a feature is based on its presence in multiple feature subsets. In this regard, the final feature subset is based on the consolidation of the base subsets. Typical ways to arrive at a final feature subset is to use set-theoretic combining techniques such as the intersection of all the base feature subsets, or union of all the subsets or multi-point intersection is also one of the favored techniques in this regard [29]. It is reported that intersection tends to produce very restrictive solutions that do not tend to produce good results [31]. While the union may result in a very large feature subset and the case of cost-based feature selection, each feature may carry a specific cost of data acquisition and hence, the overall cost of the solution may increase in this case. Other approaches for combining base feature subsets are based on classification performance of the subset [32] or data complexity measures i.e. features resulting in decreased theoretical complexity of the data must be preferred [31]. Ranking based methods leverage a number of feature scoring measures such as information gain, symmetric uncertainty, Gini index, chi-square, among others [33]. These techniques evaluate the relevance of a feature in terms of predicting the class label. Furthermore, an ordered ranking of features is returned by these base feature selection methods. Subsequently, a set of individual feature rankings are consolidated into a single ranking by taking into account the feature score and/or rank of a feature in a given list. A feature that has consistently received higher scores or is consistently placed higher in the individual scoring lists, tends to preserve its relevance in the final feature list. Since the ranking methods assign feature relevance to all the features in the dataset, therefore, a threshold value is used to select only the most relevant set of features. The threshold can be taken as a fixed value such as the top 10% of the features or it may depend on the nature of the dataset. In this study, we specifically deal with ranking approaches. It is pertinent to mention that; threshold may be applied before or after combining results of the base methods. Table 4 provides a summarized comparison of both ranking and subset based ensemble approaches.

TABLE 4. Summarized comparison of ensemble feature selection approaches.

Approach	Methodology	Merits	Demerits
Ranking-based methods	A set of feature relevance scores are used. Multiple ordered lists are obtained in the process. A final ordered list is generated through an operation such as 'min', 'max', 'arithmetic min', and 'geometric mean'. Once a final ranking is available then a threshold value is used to select a final feature subset solution is produced	• Computationally less expensive • More insightful i.e. produces ordered feature rankings for all the features	• Feature redundancy • Threshold selection is a challenging task
Subset-based methods	A set of feature subsets are produced by multiple feature selection techniques. These partial solutions are combined through set operations such as 'intersection', 'union', and 'multi-point intersection', among others	• Comparatively more accurate • Feature redundancy is addressed	• Computationally less efficient • Partial feature relevance is produced

This study focuses on the CKD diagnosis problem, therefore, most of the techniques discussed in this section are related to the feature selection and classification of the CKD problem. Polat *et al.* [13] proposed a hybrid feature selection technique for the CKD diagnosis. In this technique, the authors have demonstrated that reducing the number of features has a positive impact on the resultant accuracy of the Support Vector Machine (SVM) classifier. The subset generating mechanism is based on the Best First search while the SVM-based subset evaluator is used for assigning fitness to the candidate feature subsets. It is reported that the aforementioned feature selection approach selected the top 11 features and achieved an accuracy rate of 98.5% on the CKD dataset was acquired from the University of California at Irvine (UCI) benchmark dataset repository [34]. Almansour *et al.* [21] proposed an early stage CKD diagnosis solution in order to assist the nephrologist in decision making. The effectiveness of the proposed approach is demonstrated on the benchmark CKD dataset from the UCI. One of the objectives of this study is to provide a comparative analysis of different classification techniques such as SVM and Artificial Neural Networks (ANN). It is reported that on the CKD problem ANN achieved higher predictive accuracy than that of the SVM model. It is pertinent to mention that the dataset was apriori treated for the missing values using the mean substitution technique. This study also investigated the effects of FS on the predictive accuracy of both SVM and ANN classifiers. It is concluded that FS has a positive impact on the CKD problem. Furthermore, the top 12 features are selected in the final solution.

Ogunleye and Qing-Guo [10] proposed an ensemble decision tree-based classification approach for the CKD diagnosis problem. In this study, a number of widely used classification techniques are compared on the CKD dataset and XGBoost is subsequently selected as a base model. Furthermore, a set-theoretical rule is used for combining features from different feature selection methods such as recursive feature elimination, extra tree classifier, and univariate selection. Finally, a set of 12 features is selected. The dataset is imputed for missing values using the median substitution. The final evaluation results are reported on an out-of-sample set of about 10% data i.e. 40 cases. Qin *et al.* [12] investigated the effects of missing values substitution for the CKD diagnosis problem. In this regard, the authors have taken into account a number of classifiers such as logistic regression (LG), Random Forest (RF), SVM, k-nearest neighbor (KNN), Naïve Bayes (NB), and ANN. RF achieved the highest predictive accuracy among the aforementioned classifiers followed by the LG model. The experiments are performed on the CKD benchmark dataset from the UCI. The KNN (with $K = 11$) is selected for dataset imputation and subsequently, a set of top 11 features is selected. Furthermore, an integrated model based on Perceptron learning is used that takes probabilities from both RF and LG models. The resultant model achieved a slightly higher predictive accuracy than that of the individual models of RF and LG. It can be observed that decision tree-based ensemble modeling approaches consistently produce models with high predictive accuracy for the CKD problem as reported in [10], [12].

Wibawa *et al.* [11] investigated the CKD diagnosis problem from the cost-accuracy trade-off perspective. The authors demonstrated the efficacy of machine learning-based CKD diagnosis in terms of identifying useful features that are not taken into account by GFR estimation equations. In this study, RF is used as a base classifier, while the LASSO regularization method is used to rank the features with respect to their predictive capability. Finally, a set of 5 top most predictive features are selected that are cost-effective as well. All the experiments are performed on the UCI benchmarked CKD dataset. Chen *et al.* [29] performed an extensive study on feature subset combination methods such as union, intersection, and multi-intersection approaches. The main objective of this paper is to explore the effective approaches to combine candidate solutions from multiple feature selection methods. Three well-known feature selection methods i.e., Principal Component Analysis (PCA), Genetic Algorithms (GA), and Classification And Regression Trees (CART) are used in the experimentation. The candidate approaches are evaluated using the ANN classifier. The stock prediction problem is considered as a case study through which it is demonstrated with the intersection between PCA and GA, and the multi-intersection of PCA, GA, and CART performed comparatively better in terms of predictive accuracy of the resultant ANN model.

Ali *et al.* [30] proposed a unified ensemble feature selection approach in which the authors combine feature

TABLE 5. Summarized comparison of the key approaches to feature selection.

Ref.	Methodology	Merits	Demerits
[13]	A hybrid approach is used that is based on the wrapper and filter methods to select a final feature set. SVM is used to model the final decision for the CKD problem	The feature selection approach produces highly useful features for the SVM classifier, therefore, the final model is very accurate	Feature selection approach is based on a specific classifier, therefore, the results may not generalize to other classifiers
[21]	Two popular classification models are compared for the CKD dataset based on a set of selected features	The proposed feature selection technique is simple to implement and would incur less computational time	A large feature set is selected for the CKD problem than other comparative techniques for the same dataset
[10]	Based on a set-theoretic heuristic a set of features are selected and subsequently, an XGBoost model is built	XGBoost classifier is demonstrated to produce a highly accurate model based on the selected features	A large feature set is selected for the CKD problem than other comparative techniques for the same dataset
[12]	LG and RF are used in an ensemble configuration where the Perceptron model is finally built on the basis of a selected set of features	Ensemble modeling approach based on Perceptron learning produces a highly accurate model	Final feature subsets are selected in a post-hoc manner based on only the RF feature weight-age.
[22]	Feature subset is selected based on the LASSO method while RF is subsequently used as a classification model	High accuracy and low solution cost are reported on the basis of a set of features selected by the decision-maker	Cost is considered in a post-hoc manner, where no automated approach is proposed for the final feature set selection
[29]	Multiple feature scoring methods are used for obtaining feature relevance scores, finally, an ANN is used for modeling the stock prediction	Multiple candidate solution approaches are demonstrated for the stock prediction problem	GA generally incurs a high computational cost, while CART is only used for feature weight-age, therefore, also incur an additional cost element in the final solution modeling
[30]	Multiple filter methods are used for obtaining a unified feature ranking and a final feature subset is produced based on a fixed threshold value	Filter methods are combined using simple heuristics	A fixed threshold value is used in the study i.e. 45% of the feature set
[28]	Candidate feature subsets are selected in an ensemble configuration, subsequently, a decision tree model is built based on the selected features	Filter methods are combined using simple heuristics	A fixed threshold value is used in the study i.e. 1/3rd of the features from individual techniques in the ensemble

weight-age techniques and generate a combined feature weight-age. Furthermore, based on extensive experimentation on both medical and non-medical datasets, it is reported that the top 45% features of any dataset generally provide good enough accuracy. Hence, a fixed threshold is proposed. The feature weight-age methodology is evaluated based on a number of classifiers i.e., NB, C4.5, KNN, RIPPER, and SVM. The individual feature scoring technique in the ensemble include information gain, chi-squared, gain ratio, symmetric uncertainty, and signification. Osanaiye *et al.* [28] proposed an ensemble-based feature selection technique. The proposed technique is based on a set of filter methods such as information gain, gain ratio, chi-square, and ReliefF to comprehensively assign a score to the features. In this study, the intrusion detection problem is taken as a case study to demonstrate the effectiveness of the proposed approach. A set of candidate feature sets are selected from each feature scoring method using the top $1/3^{rd}$ features in the dataset. Hence, four subsets are obtained. A final feature set is selected based on the intersection of the candidate sets. The decision tree classifier is used to evaluate the effectiveness of the generated solution. Table 5 provides a comparative analysis of some of the key feature selection techniques considered in this study.

The aforementioned studies applied feature selection techniques to the structured data. The image data is yet another data modality where deep learning techniques have been successfully applied yielding highly successful results such as [35], [36]. Wen *et al.* [35] proposed a multilabel image classification approach that used the co-projection of features and labels present in the dataset. The gist of the method is to project both labels and image features to a common latent vector space. In this way, the frequently occurring features and labels do appear closer in the latent space as well. In another study, Li *et al.* [36] proposed a domain adaptation approach for object detection in medical images. As the medical images are expensive to obtain, resulting in insufficient records in the training datasets. In this regard, the domain adaptation models are one of the appealing alternatives to obtain accurate results regardless of the differences in the data distribution.

III. METHODOLOGY

This section deals with a detailed description of our proposed methodology. As mentioned in Section I, the main impetus of the methodology is to deal with a group of features in a cost-sensitive manner. Most of the ranking techniques assume that cost of feature acquisition is independent of other features i.e. cost is not shared among different features. Therefore, each feature is evaluated independently based on its predictive score and associated cost. Although the aforementioned

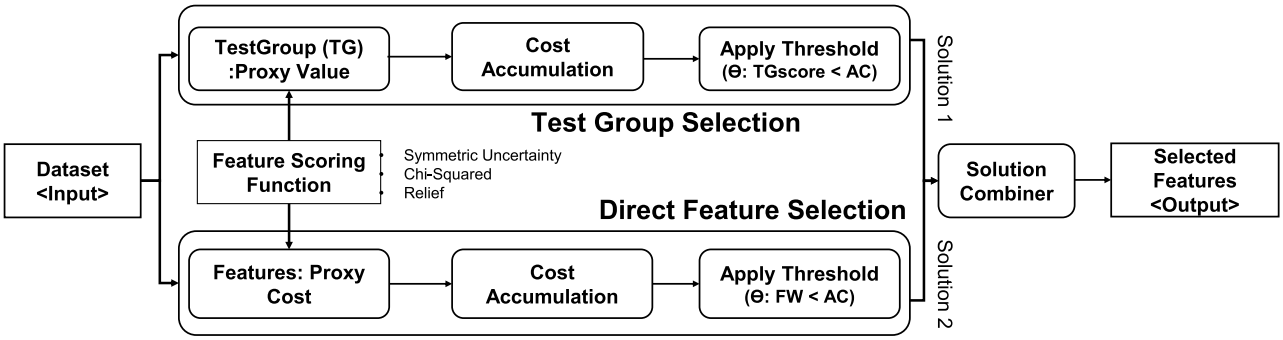

FIGURE 4. Schematic illustration of the proposed hybrid approach.

problem formulation simplifies the selection of a set of features, it does not account for situations where features are dependent on other factors such as medical tests e.g. urine test [2], [3]. In medical diagnosis applications, acquiring data for a set of features is dependent on carrying out certain medical tests. Hence, a relatively small set of medical tests may cover all of the features present in the dataset, where the cost is distributed among features. In this regard, a medical test is comprised of a set of features whereas multiple features are grouped under a certain test. So the challenge is to assign a proxy merit value to the medical test based on the combined worth of features. And reciprocally, assign a proxy cost to individual features based on the overall cost of the medical test. We address these problems in parallel. In the first approach, the cost factor is given precedence over the feature weight-age. In this case, a subset of cost-effective test groups (TGs) are selected i.e. each medical test is assigned a relevance score, and based on the cost-aware threshold value a subset of tests is selected. In the second approach, features are directly selected based on their collective ranks i.e. each feature is assigned a relevance score based on its predictive power, and subsequently a cost-aware threshold value is selected. It is pertinent to note that in both the aforementioned techniques, features are considered in groups. In this research, we have used the heuristic of intersection between the score and accumulated cost curves [21] for selecting a threshold value as shown in Figure 6. Both solution-1 (S1) and solution-2 (S2) produces subset of features that are later combined into a consolidated solution. The partial solutions obtained from the aforementioned techniques have different characteristics; the solution obtained from S2 tends to minimize the collective cost of features regardless of the cardinality of the solution set, while the S2's solution tends to enhance the collective predictive value of the solution while optimizing the solution set. Figure 4 depicts the schematic representation of the proposed ensemble technique.

A. MISSING VALUES IMPUTATION

Real-world datasets are generally not of high quality. Therefore, data preprocessing is a prerequisite and inherently an implicit step in most of the data-driven systems. Although several preprocessing steps can be applied to the selected CKD dataset such as outlier removal, data normalization, data discretization, removal of nominal features with higher stability, id-ness in numerical features, etc., we decided to apply only the data imputation to preserve the original data semantics and data size. The KNN approach is used to impute missing values where the $K = 3$. The assumption employed in imputation operation is that similar instances would have similar characteristics. Hence, the target instance having one or more missing values can be treated with the local information obtained from neighboring instances. Please note that in this study we apply KNN after each imputation to ensure the validity of the information obtained from the selected sub-sample. Numeric attributes are treated with median values while nominal attributes are imputed with mode values. Since, outliers are not explicitly removed from the dataset therefore, imputation using median values can reduce the susceptibility to the outliers. It is observed that around 60% of the records in the CKD dataset contained missing values. Mixed Euclidian measure is used for calculating the similarity among different instances. Moreover, the median value is used for the imputation of numeric attributes while nominal attributes are imputed using the mode value.

B. FEATURE SCORING FUNCTION

Feature scoring is one of the key steps in the proposed methodology. Therefore, it is of utmost importance that the features are not scored based on any spurious patterns. Figure 5 illustrates a schematic representation of a generic feature scoring function. Furthermore, as it is also reported in the literature that the filter-measures tend to produce results that are more general than the wrapper-methods [37]. Therefore, based on this observation we have employed a set of filter measures in the proposed methodology. Moreover, the individual feature weight-ages obtained from these measures are used for obtaining ranked feature lists. Several feature weight-age techniques are reported in the literature that is based on information-theoretic measures, distance measures, consistency measures, and correlation measures among others [30], [33], [37]. One important factor in the

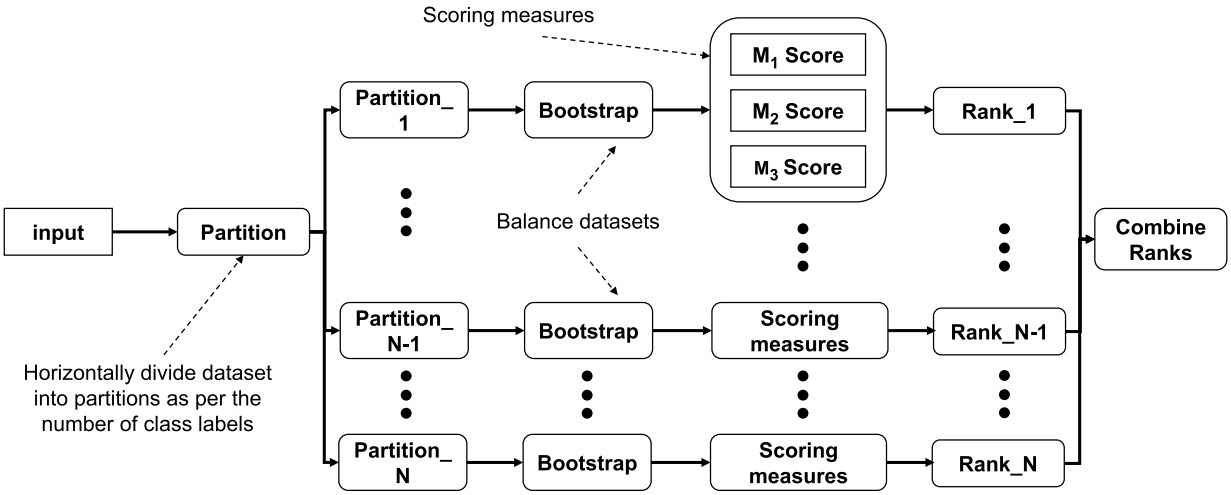

FIGURE 5. Schematic illustration of the generic feature scoring function based on multiple partitions.

selection of a set of measures is the diversity among the produced candidate solutions. So long as the measures in an ensemble are complementary to each other, the obtained feature ranking would be more comprehensive. In order to find the importance of a feature in a given dataset, we have used symmetric uncertainty, chi-square statistic, and a multivariate method called Relief [38]. A large number of feature selection methods use information-theoretic measures for univariate feature selection. In this regard, information gain is one of the popular techniques. It is reported that information gain is biased towards attributes having a large number of distinct values as the data partitions obtained for each distinct value would be having a homogeneous class distribution. Therefore, symmetric uncertainty is used as one of the feature scoring techniques that has a low bias towards estimation. A similar low bias is observed in the case of Relief as well [39]. It is also reported that both these measures exhibit a linear growth as a function of the values present in a feature, with steepness proportional to the number of classes [39]. Chi-square is another popular statistical measure that is applied to test the independence of two events. In feature selection, these two events can be in terms of the occurrence of the feature and the occurrence of the class. The symmetric uncertainty is a widely used variant of normalized mutual information. The application of symmetric uncertainty in feature scoring is in terms of information exchange between two feature vectors. In a univariate case, one of the vectors is an independent feature such as the *age* of a patient, while the other vector is the dependent variable i.e. class variable such as *diagnosis*. This measure quantifies the mutual dependence of two variables as given in Eq. 1.

$$SU(A,B) = 2[\frac{MI(A,B)}{H(A)+H(B)}] \tag{1}$$

where $MI(A,B)$ is the mutual information between feature A and feature B, and entropy of features A and B is computed by $H(A)$and $H(B)$, respectively. The chi-square is used to compare expectations with that of original observed data. In feature selection, this test is used to evaluate the nature of the relationship between two variables. Using observed and counted statistics one can test the independence of whether a strong correlation exists between an independent variable and a dependent variable or not. Chi-square is computed as given in Eq. 2.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

where O_i denotes observed values for an instance i, and E_i represents expected values. The third ranker approach is based on the Relief algorithm [8], which provides a feature score based on their interactions, and thereafter the provided scores can be subsequently used for generating features ranks. Relief algorithm tends to compute a feature vector W according to Eq. 3.

$$W_i = W_i - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2 \tag{3}$$

where *nearHit* refers to the closet same-class instances, *nearMiss* refers to closet instances from other classes, and $(x_i - nearHit_i)$ denotes the Euclidean distance between the two points. Please note that the combined ranks are normalized between 0 and 1.

In order to generate reliable feature scores, the dataset is horizontally partitioned into multiple non-overlapping subsets that are balanced through bootstrap with replacement technique, as depicted in Figure 5. A set of filter measures is applied on each bootstrapped dataset and feature ranks are obtained. Afterward, all the ranks are consolidated. The main objective of creating multiple partitions and repeatedly applying the scoring measures is to obtain unbiased, diversified, and stable feature ranking. It is one of the key steps in the proposed approach. As a rule of thumb, a dataset is divided into a number of partitions that are equal to the number of classes in the dataset. Subsequently, a set of scoring measures are applied to the bootstrapped dataset.

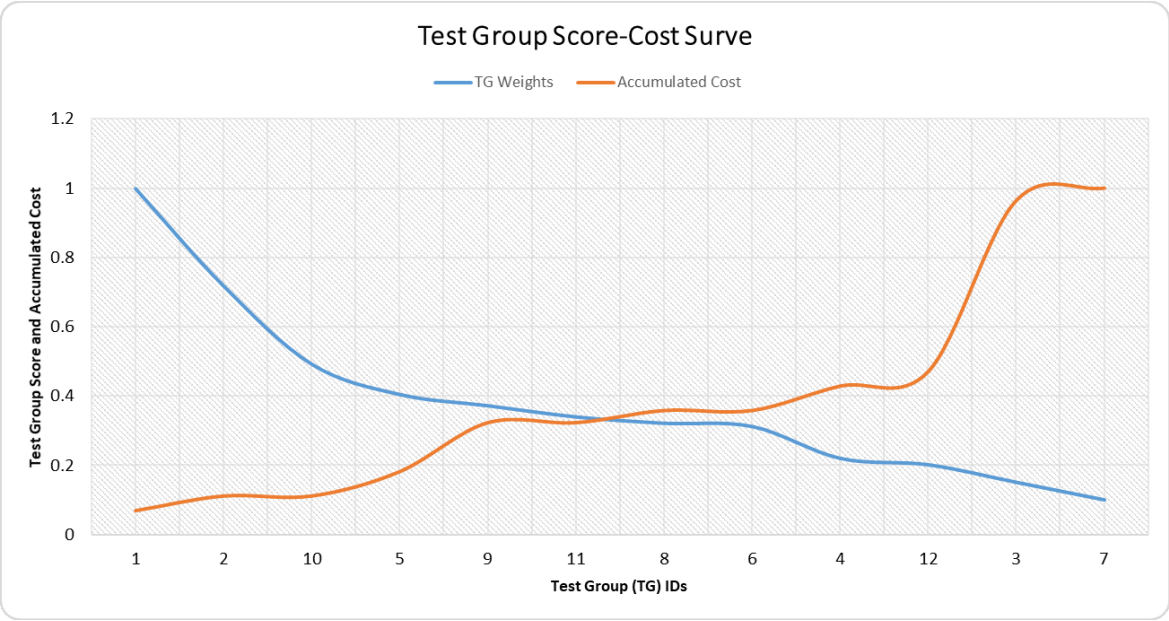

FIGURE 6. Solution-1 Approach based on Test Group (TG) and Accumulated Cost curve.

C. SOLUTION-1: FEATURE SELECTION BASED ON THE TEST GROUP

Once features are scored based on their relative ranks, then the subsequent task to find an appropriate threshold value for selecting a subset of features that are both highly predictive of the class label and are cost-effective at the same time. S1 is an indirect method in which cost-effective test groups are selected. As each test group has one or more features, therefore, all the features which have a membership with the selected test groups are selected as the final solution. Each test group is assigned a cost value and score value. Cost values for different test groups are provided in Table 8, whereas, the test group score is estimated based on the global ranks of the features in a particular test group as given in Eq. 4. In the case of the individual feature relevance score assignment, each feature is globally assigned a relevance score based on its predictive score among all the features. Subsequently, all the features are arranged into an ordered list with respect to their relevance score. In order to assign relevance to a TG, first all the features are divided among different TGs. The score of each TG is based on the summation of ranks of all the features in a TG to the collective ranks of all the features in the dataset.

$$ScoreTG(F) = \frac{\sum_1^j \hat{f}_j}{\sum_1^k f_k} \tag{4}$$

where j is defined as $\{j|TG(f_j)\}, f \in TG$, both$(\hat{f}, f) \in F$, and k is the cardinality of the original feature set. The score of a particular TG is the accumulated rank of all the features that belong to the TG over the summation of all the feature ranks. Figure 6 depicts the score of TGs, where TGs are sorted with respect to their relevance score.

The blue line in Figure 6 denotes the worth of a particular test group while the orange line shows the accumulated cost of the test groups. Please note that both the score and the cost values are normalized. The threshold selection is based on the point of intersection between these two curves.

The accumulated cost value for the sorted test groups, TG, is calculated using Eq. 5.

$$\begin{aligned} TG_{Cost}(e_i) = {} & GCv(e_i) + GCv(e_{i-1}) \\ & + GCv(e_{i-2}) + GCv(e_{i-3}) \\ & + \ldots + GCv(e_0) \end{aligned} \tag{5}$$

where e denotes one specific test group, $i = 1, 2, 3, \ldots, m$ i.e. m is the total number of test groups in GCv, and GCv represents actual cost values for all the test groups. Eq. 5 represents an accumulation cost function at any given point i where $GCv(e_0) = 0$. It is important to note that the GCv is already in sort order according to the Test Group relevance (refer to Algorithm 1, line 7).

D. SOLUTION-2: DIRECT FEATURE SELECTION

In the second approach, features are directly selected based on their score and a proxy cost value. As it is mentioned that the cost is directly associated with a test group, therefore, each feature is assigned a cost value based on its relative rank in a particular test group i.e. higher the rank, the lower would be the cost. The main objective of this approach is to select highly predictive features while at the same time avoid the selection of lower score features in a particular test group. Unlike the earlier approach in which all the features in a selected test group are part of the final solution, in this approach features are directly selected regardless of their membership with any particular test group. Features having

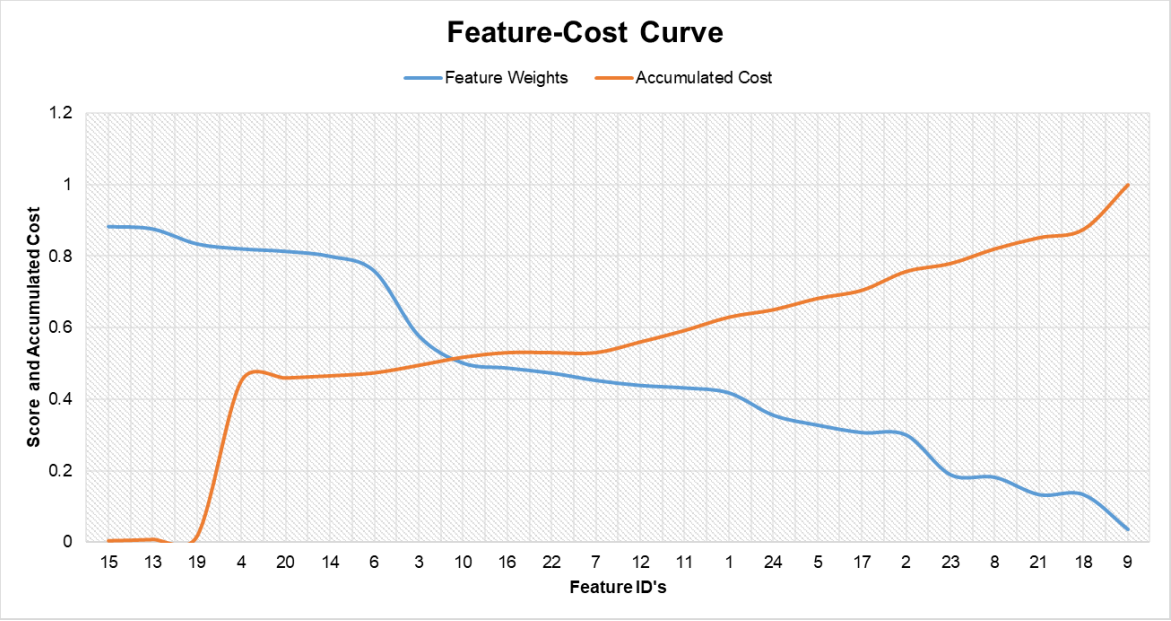

FIGURE 7. Pictorial representation for the direct and indirect group based feature selection approaches (*DFS-CT*).

Algorithm 1 Solution-1 (Indirect Feature Selection)

Input: Dataset D, Group Cost vector GCv, Filter Measures M

Output: Final Solution Set S

1: **for** measure m in M **do**:
2: Feature_list = Score_feature(D, m)
3: **end for**
4: TestGroup_list = AssignTGSCore(*Feature_list*) using Eq. 4
5: TG = Sort(*TestGroup_list*)
6: **for** element e in TG **do**
7: $TG_{Cost} = GCv(e)$ using Eq. 5
8: **end for**
9: T = Intersection (TG, TG_{Cost})
10: Selected_TG = Selected (TG, T)
11: S = Unpack(*Selected_TG*)
12: return S

a high score and high cost have a relatively higher chance of selection in the direct feature selection as compared to the indirect feature selection i.e. S1.

In case of cost assignment, a feature's TG is taken into consideration. As cost is directly associated with TG, therefore, features that belong to a particular TG share the cost factor. In this regard, the cost can be distributed uniformly among all the features of a particular TG, given by Eq. 6. In this case, TG is comprised of a number of features. Therefore, the cardinality of TG is based on the number of member features in the TG. Furthermore, the cost of TG is externally provided by the decision-maker. The feature cost can also be assigned based on the relative rank of a feature in the TG i.e. proportional assignment, higher the relative rank lower is the cost as given in Eq. 7.

$$Cost(f) = \frac{Cost(TG)}{|TG|} \tag{6}$$

$$Cost(f) = Cost(TG) * RelativeRank(f) \tag{7}$$

The $RelativeRank(.)$ of a feature is based on a features local rank among divided by the summation of all the ranks in a given TG. It is observed that the overall difference between both the aforementioned cost assignment techniques is not significant. In this study, the proportional cost assignment is used. As can be seen in Figure 7, each feature is associated with a score as well as a cost factor. The feature score, blue line, is based on the average feature weight-age across multiple feature scoring measures while the cost, orange line, is based on the relative cost of the feature in a particular test group. Similarly, the threshold value is based on the point of intersection between the feature score and feature cost curves. The direct feature selection can be performed in two ways. A combined feature score can be obtained by taking an average over different feature weights as depicted in Figure 8. We call this approach direct feature selection - combine then threshold (*DFS-CT*). In this case, the threshold value is based on the combined worth of a feature which is in turn computed through multiple feature scoring measures. As a list of features is obtained, therefore, a threshold value is required to select a subset of features.

The second approach, direct feature selection - threshold then combine (*DFS-TC*), for computing direct feature selection is to apply the threshold value to the individual measures, in parallel. In this case, each measure produces

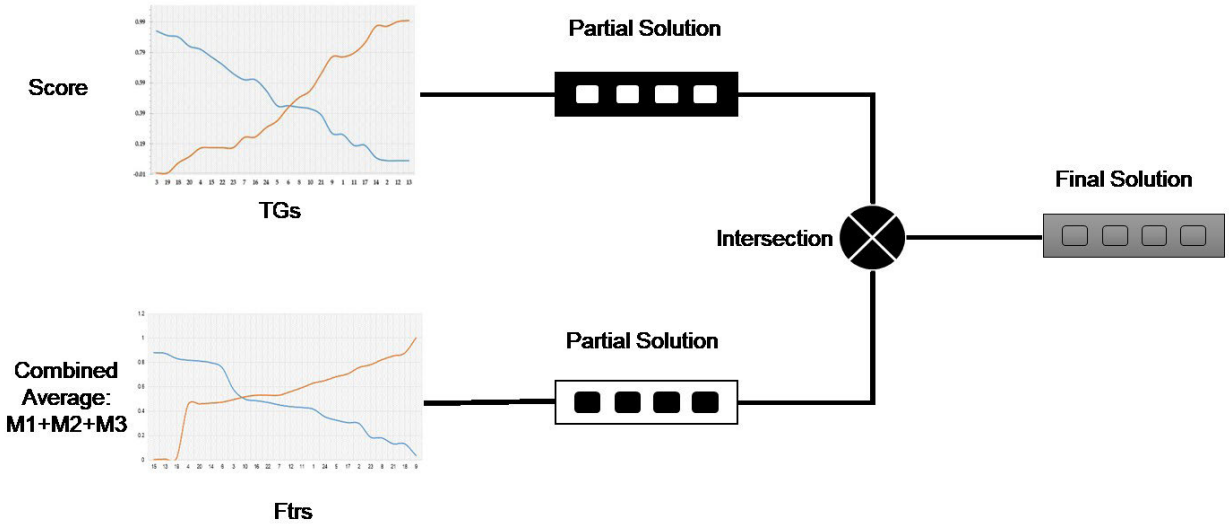

FIGURE 8. Abstract flow of User-Centric Adaptive Intervention methodology.

Algorithm 2 Solution-2a - Direct Feature Selection - Combine Then Threshold (*DFS-CT*)

Input: Dataset D, Group Cost vector FCv, Filter Measures M

Output: Final Solution Set S

for measure m in M **do**:
    Feature_list = Score_feature(D, m)
end for
F_list = Sort(*Feature_list*)
for element e in F_list **do**
    $Feature_{Cost} = FCv(e)$ using Eq. 8
end for
T = Intersection (F_list, $Feature_{Cost}$)
S = Selected(*F_list, T*)
return S

Algorithm 3 Solution-2b - Direct Feature Selection - Threshold Then Combine (*DFS-TC*)

Input: Dataset D, Group Cost vector FCv, Filter Measures M

Output: Final Solution Set S

for measure m in M **do**:
    Feature_list[m] = Score_feature(D, m)
    F_list[m] = Sort(*Feature_list[m]*)
    for f in F_list[m] **do**
        $F\_List[m]_{Cost}$ = FCv(f) using Eq. 8
    end for
    T = Intersection(*F_list[m], Cost-F_list[m]*)
    Selected_Features = Selected (*F_List, T*)
    for element e in Selected_Features **do**
        $Feature_{Cost} = FCv(e)$ using Eq. 8
    end for
    T = Intersection($F\_list[m]_{Cost}$, $Feature_{Cost}$)
    S[m] = Selected(*F_list[m], T*)
end for
S = Combine(*s[M]*)
return S

a candidate feature subset which is subsequently combined. Unlike *DFS-TC* where the same size ranks are consolidated, in this case, the size of the candidate feature subsets may not be the same. Different subset combining techniques can be used such as union, intersection, and multi-point intersection. The union results in a large number of features while the intersection is the most restrictive of the three. The multi-point intersection tends to reflect the majority voting scheme and hence it has characteristics of both the aforementioned techniques. Therefore, we multi-point intersection is favored in the case of *DFS-TC*. Figure 9 depicts the *DFS-TC* approach.

Algorithm 2 denotes the process of feature selection through the *DFS-CT* approach, while *DFS-TC* is represented by Algorithm 3. The major difference in both these approaches is in terms of the application of threshold operation. Both the aforementioned approaches for direction feature selection use the same accumulated cost formula as given in Eq. 8.

$$\begin{aligned} Feature_{Cost}(e_i) = {} & FCv(e_i) + FVc(e_{i-1}) \\ & + FCv(e_{i-2}) + FCv(e_{i-3}) \\ & + \ldots + FCv(e_0) \end{aligned} \tag{8}$$

where e denotes a feature, i $= 1, 2, 3, \ldots,$ n i.e. n is the total number of features in FCv, and FCv represents actual cost values for all the features based on their membership in different test groups. Eq. 8 represents an accumulation cost function at any given point i where $FCv(e_0) = 0$. Table 5 provides a summary of the functions used in both Algorithm 1 and Algorithm 2.

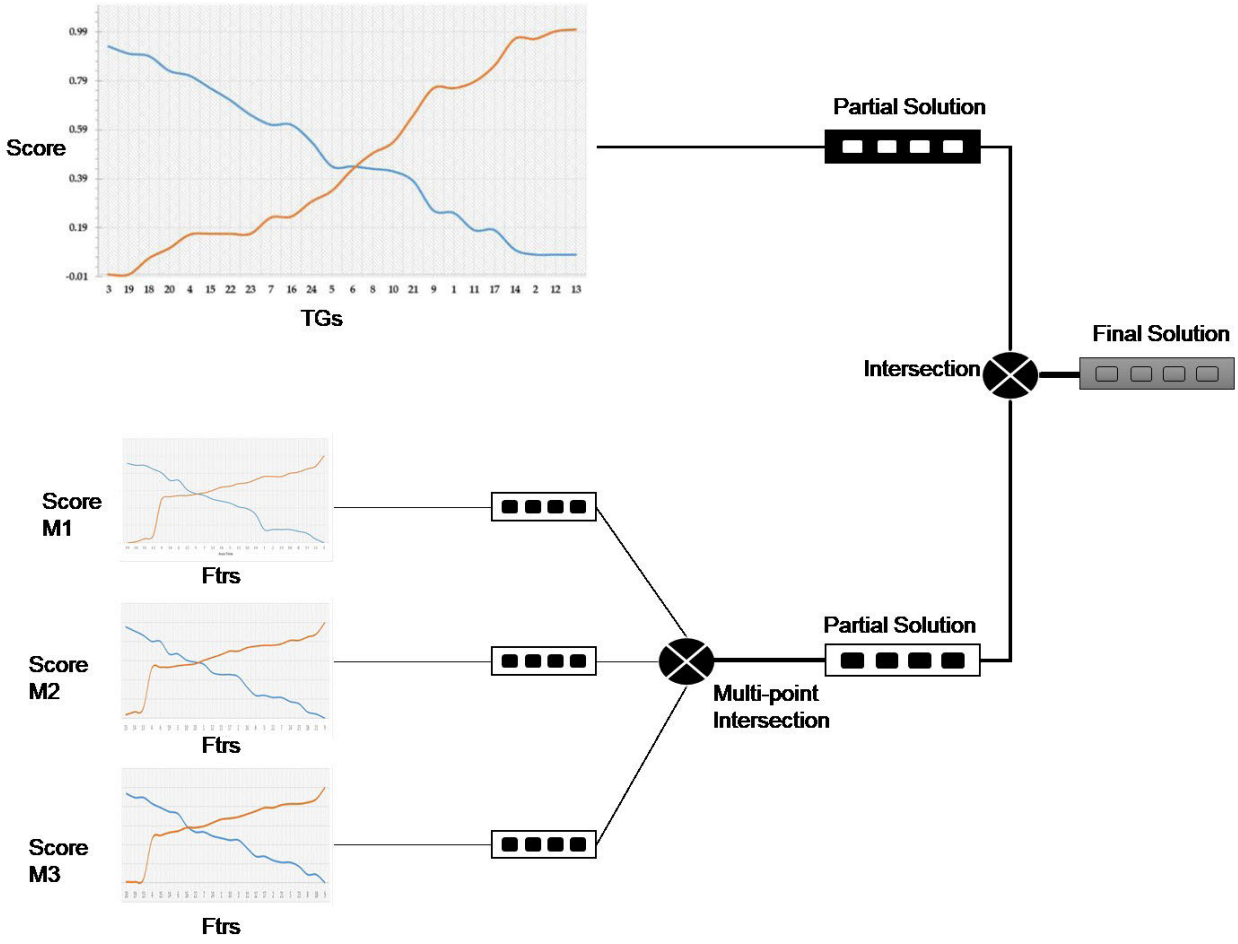

FIGURE 9. Pictorial representation for the direct and indirect group based feature selection approaches (*DFS-TC*).

TABLE 6. Explanation of functions used in the proposed Algorithm 1 and Algorithm 1.

Function Name	Input	Output	Purpose
Score_feature()	D: Dataset m: Univariate Measure	List: Feature relevance (FR) scores w.r.t 'm'	Assign each feature a relevance score
AssignTGScore()	List: Feature list (FR)	List: Test Group (TG) relevance scores w.r.t 'Feature	Assign each test group a relevance score based on its member features' scores (FR)
Sort()	List: Feature or TG list	List: Ordered list of features or test groups	Rank features/TGs based on their relevance score
FCv or GCv()	Element: a specific feature or TG	Value: assign cost value to a feature/TG element based on its position in the list	Compute the accumulated cost of an element in the list i.e. feature/TG ordered list
Intersection()	List: Relevance list (F/TG), Accumulated Cost list (F/TG)	Value: first interaction point between the relevance list and accumulated cost list	The interaction point serves as a threshold value
Selected()	List: FR Value: Threshold value 't'	List: Solution S, a set of features are retained that are above 't'	To select a subset of features or TGs
Unpack()	List: S in case of the TG	List: all the features that are a member of selected TGs	A final solution is obtained in case of TG selection approach
Combine()	Lists: Partial lists corresponding to candidate feature subsets	List: a consolidated list is obtained through multi-point interaction operation	Purpose: to combine multiple candidate solutions into a final solution

E. SOLUTION COMBINER

The partial solutions obtained from S1 and S2 are combined using set-theoretic interaction operation. S1 is geared towards a low-cost solution while S2 is designed to obtain a set of highly predictive features with less regard to the overall cost of the solution. Both S1 and S2 are designed to achieve two different objectives. Therefore, the combiner operation is performed to form a solution that reflects the characteristics of both S1 and S2. Features that are present in both the partial solutions are selected in the final set. It is important to note as

TABLE 7. Chronic Kidney Disease (CKD) dataset characteristics along with test group (TG) association.

ID	Attribute	Test Group	Description	ID	Attribute	Test Group	Description
1	Age <*age: numerical*>	12	In years	13	Sodium <sod: numerical>	5	mEq/L
2	Blood Pressure <*bp: numerical*>	6	Mm/Hg	14	Potassium <*pot: numerical*>	5	mEq/L
3	Specific Gravity <*sg: numerical* >	1	1.005, 1.010, 1.015, 1.020, 1.025	15	Hemoglobin <*hemo: numerical*>	2	Gms
4	Albumin <*al: numerical*>	3	0–5	16	Packed Cell Volume <*pcv: numerical*>	2	Integer valued
5	Sugar <*su: categorical*>	7	0–5	17	White Blood Cells Count <*wc: numerical*>	2	cells/cumm
6	Red Blood Cells <*rbc: categorical*>	1	1: Normal, 0: Abnormal	18	Red Blood Cells Count <*rc: numerical*>	2	millions/cmm
7	Pus Cell <*pc: categorical*>	1	1: Normal, 0: Abnormal	19	Hypertension <*htn: categorical*>	6	1: Yes, 0: No
8	Pus Cell Clumps <*pcc: categorical*>	1	1: Present, 0: Absent	20	Diabetes Mellitus <*dm: categorical*>	7	1: Yes, 0:No
9	Bacteria <*ba: categorical*>	9	1: Present, 0: Absent	21	Coronary Artery Disease <*cad: categorical*>	8	1: Yes, 0: No
10	Blood Glucose Random <*bgr: numerical*>	7	mgs/dl	22	Appetite <*appet: categorical*>	10	1: Good, 0: Poor
11	Blood Urea <*bu: numerical*>	4	mgs/dl	23	Pedal Edema <*pe: categorical*>	11	1: Yes, 0: No
12	Serum Creatinine <*sc: numerical*>	4	mgs/dl	24	Anemia <*ane: categorical*>	2	1: Yes, 0: No

there are two alternative approaches for S2 i.e. *DFS-CT* (refer to Figure 8) and *DFS-TC* (refer to Figure 9), therefore, any one of these two may be selected while the *combine* operation would remain the same.

IV. EXPERIMENTATION AND RESULTS

This section deals with the experimental details of the CKD diagnosis case study. A summarized description of the CKD dataset is provided in this section along with its overall quality. The experimentation design is divided into a number of steps such as first we evaluate the diversity of the ensemble elements i.e. feature scoring measures such as symmetric uncertainty, chi-square, and Relief. Then we evaluate the baseline results on the dataset without using any feature selection technique. In this regard, 7 classifiers are used to evaluate the comparative methods in a comprehensive manner. Afterwards the efficiency of both solution-1 and solution-2 is demonstrated in the selection cost-effective feature subset. As solution 2 can be configured in multiple ways, therefore, the best configuration for the CKD dataset is selected. The final evaluation also includes a student t-test to compare the statistical difference between the proposed and other comparative techniques.

A. DATASET DESCRIPTION

The dataset used in this case study is taken from the online benchmark repository of the University of California [34]. It is a real-world dataset of CKD patients prepared by Apollo Hospitals, Tamilnadu, India over two months. Furthermore, the cost factor associated with the dataset is acquired from the Pakistan Institute of Medical Science, Islamabad, Pakistan. It is pertinent to note that both the dataset and the data acquisition cost are included in this case study from a South Asian perspective. The CKD dataset contains information about 400 patients. In this regard, a wide variant of information is available for each patient such as a patient's age, blood pressure level, specific gravity, red blood cells, blood urea, diabetes mellitus, anemia, etc. In total each patient case is characterized by 24 different features. This data set contains both nominal as well as numeric variables. The final decision reflects whether a given patient has CKD disease or not, therefore, a binary variable is used to model the decision. Furthermore, the dataset is not heavily skewed towards any particular class i.e. it contains 250 positive patients and 150 negative patients, respectively. The dataset is treated for missing values using median value imputation. The dataset is divided into multiple test groups (TG) where one or many features mutually exclusively belong to each TG. Furthermore, each TG is associated with a cost factor. For example, urine analysis is a TG that is composed of 4 features i.e. specific gravity, pus cells, red blood cells, and puss cells clumps. The cost of this test is 100 PKR. Likewise, all the features are divided into 12 groups. Please note that data for some of the features in the CKD dataset may be acquired without conducting any specific medical test e.g. age, blood pressure reading, clinical history-related questions e.g. hypertension, pedal anemia, etc. Therefore, those features that may not require any specific test are nevertheless assigned to a TG having cost factor 0 PKR. Table 7 provides details on the key characteristics of the CKD dataset and Table 7 shows test groups and their respective costs.

In this research, the cost values for different TGs are taken from a public hospital that is heavily subsidized by the government. Therefore, private laboratories may charge differently for the tests mentioned in Table 8. Hence, it is pertinent to note that the feature selection solution for a public setup may differ from that of a private setup due to their difference in the incurred cost of conducting medical tests.

TABLE 8. Test groups and their respective cost in Pakistani Rupees (*PKR*).

Test Group	Pakistani Rupees (*PKR*)	Test Group	Pakistani Rupees (*PKR*)
1	100	7	50
2	60	8	50
3	700	9	200
4	100	10	0
5	100	11	0
6	0	12	60

According to the cost values mentioned in Table 8, the total cost for all the tests is 1420 PKR. The total cost serves as a baseline value to compare the cost-effectiveness of the comparative feature selection approaches.

B. EXPERIMENTAL SETUP

In order to evaluate the proposed approach over the CKD case study, a set of 7 classifiers are used namely as Naïve Bayes, Logistic Regression, Artificial Neural Networks, Classification And Regression Trees, Random Forest, Gradient Boosted Trees (GBT), and Support Vector Machine. All the simulation is performed on AMD Ryzen 3 200 G processor with 8 GB RAM, and 64-bit Windows 10 Enterprise Edition. Furthermore, RapidMiner Studio 9.6 is used for simulation [40]. Table 9 provides parameters used for the classification models.

TABLE 9. Classification models parameters.

Classifiers	Parameters
Naïve Bayes	N/A
Logistic Regression	N/A
Artificial Neural Network	Layers: 4 Hidden Layer size: 50 each Activation: Rectifier, Softmax
Decision Tree	Impurity measure: Gini index Maximal depth: 4
Random Forest	Number of trees: 20 Maximal depth: 7
Gradient Boosted Trees	Number of trees 20 Maximal depth: 7 Learning rate: 0.100
Support Vector Machine	Gamma: 0 C: 10

The proposed approach is compared with existing approaches through several evaluation metrics such as F1-measure and Area Under Receiver Operating Characteristics curve (AUC). First, we generated a confusion matrix to determine true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) cases. Based on these values we compute precision and recall which in turn is used for calculating F1-measure. All the experiments are performed using 5-fold cross-validation. In this case, the original dataset is divided into 5 partitions. We generate a feature selection solution using data from 4 partitions while the 5^{th} partition is used for calculating F1-measure values and AUC. This process is iterated 5 times in total, each time a different testing partition is used. The reported results are averaged over 5 different partitions. The evaluation metrics are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TN}{TN + FP} \tag{12}$$

$$F1 - measure = 2 * \frac{(Recall * Precsision)}{(Recall + Precision)} \tag{13}$$

C. RESULTS AND ANALYSIS

In this study, the ensemble for cost-sensitive feature selection is composed of three filter measures i.e. SU, chi-square, and Relief. Kendall rank correlation [41], is used to measure the pair-wise correlation of ensemble elements as shown in Table 10. The top 10 features are evaluated for each of the aforementioned measures. In this regard, as it can be seen that the correlation is closer to 0. Therefore, the null hypothesis of mutual independence is supported. As the produced partial lists are statistically different from each other, therefore it can be concluded that the produced lists are not redundant.

TABLE 10. Kendall rank correlation.

	SU	Chi-Square	Relief
SU	1	0.02899	0.094203
Chi-Square	0.02899	1	0.152174
Relief	0.094203	0.152174	1

1) BASELINE EVALUATION

In order to evaluate the efficacy of the proposed approach, we have established the baseline results where the original dataset is evaluated over 7 classifiers. No feature selection is applied to the dataset, while the missing values are treated using the median technique. The baseline results are reported in Table 11.

2) SOLUTION-1 (TEST GROUP SELECTION)

In solution-1 feature selection is based on the test group selection. In this case, each TG is assigned a score that is based on the collective rank of the features for a given TG. Figure 10 depicts the TG score, accumulated cost, and the point of intersection that serves as a threshold value for TG selection.

It can be seen in Figure 6 the point of intersection is around TG 11. Therefore, all the TGs below the threshold value are discarded. All the features are selected from the remaining *TGs* i.e. $TG = (1, 2, 10, 5, 9, 11)$. Table 12 provides detailed results for the solution-1 approach.

The total number of features selected in solution-1 is 14 and the overall cost of solution-1 is 450 *PKR*. It can be seen

TABLE 11. Baseline results.

Model	Accuracy	Precision	Recall	Specificity	F-measure	AUC	Cost (*PKR*)
NB	62.3 ± 2.0%	62.3 ± 2.0%	100.0 ± 0.0%	0.0 ± 0.0%	76.6 ± 1.5%	90.8 ± 9.0	
LR	84.3 ± 6.6%	83.2 ± 5.8%	94.3 ± 9.3%	66.9 ± 14.0%	88.2 ± 5.6%	95.2 ± 5.0	
ANN	89.5 ± 2.3%	85.4 ± 2.8%	100.0 ± 0.0%	72.8 ± 5.7%	92.1 ± 1.6%	100.0 ± 0.0	
CART	87.7 ± 4.8%	97.3 ± 3.7%	90.4 ± 10.5%	95.6 ± 6.1%	93.3 ± 4.5%	96.6 ± 3.0	1420
RF	89.5 ± 3.5%	83.9 ± 3.6%	100.0 ± 0.0%	66.4 ± 11%	91.2 ± 2.2%	99.8 ± 0.4	
GBT	73.8 ± 5.8%	86.2 ± 6.9%	100.0 ± 0.0%	71.4 ± 15.6%	92.5 ± 4.0%	100.0 ± 0.0	
SVM	92.2 ± 9.0%	71.8 ± 8.1%	97.2 ± 3.8%	35.8 ± 20.7%	82.4 ± 5.6%	84.4 ± 12.0	
Average	82.75 ± 4.8%	81.44 ± 4.7%	97.41 ± 3.3%	58.41 ± 10.4%	88.07 ± 3.5%	95.2 ± 4.0	

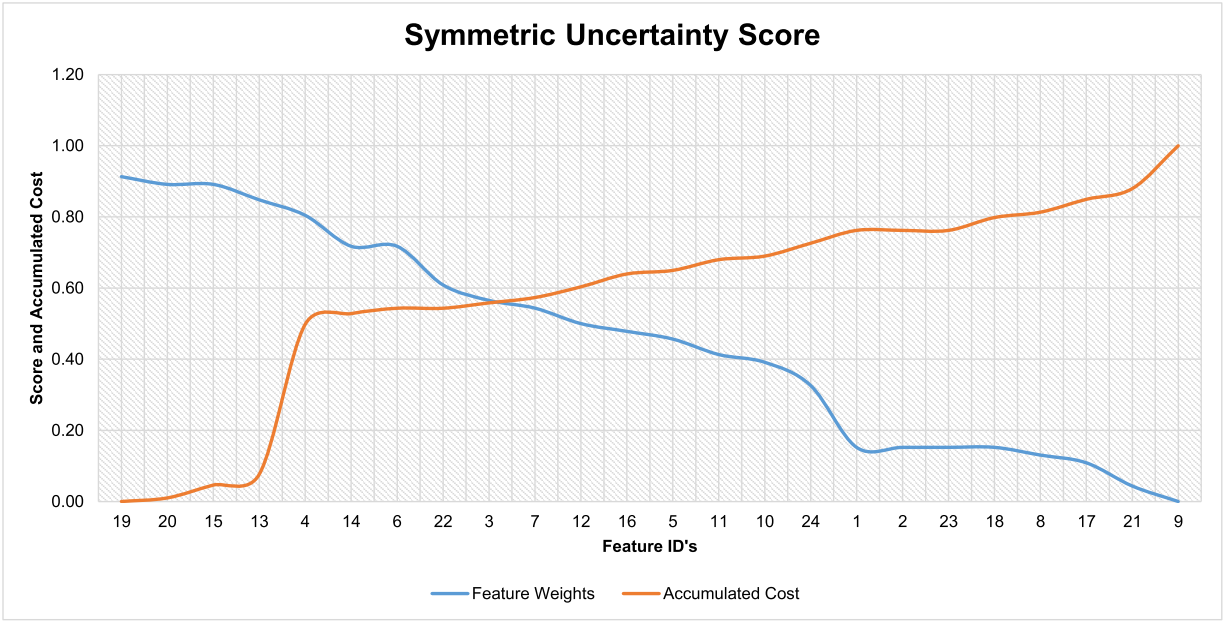

FIGURE 10. Feature-Cost curve for the Symmetric Uncertainty measure.

TABLE 12. Test group selection results (solution-1).

Model	Accuracy	Precision	Recall	Specificity	F-measure	AUC
NB	64.90±0.70	64.3±0.90	100.0±0.0	4.40±6.10	78.20±0.60	99.0±1.60
LR	95.20±0.0	64.3±0.90	100.0±0.0	6.70±6.10	78.20±0.60	83.20±9.30
ANN	98.30±2.40	98.8±2.80	98.70±3.0	97.50±5.60	98.70±1.80	100.0±0.0
CART	92.10±3.70	96.0±3.70	91.50±9.30	93.10±6.40	93.30±3.50	95.60±4.60
RF	96.50±2.0	97.40±3.50	97.20±3.80	95.30±6.50	97.20±1.60	99.80±0.40
GBT	90.40±3.60	88.0±5.10	98.60±3.20	76.10±11.90	92.90±2.50	99.0±0.90
SVM	85.10±4.0	81.60±3.80	98.60±3.20	62.50±10.70	89.20±2.70	92.20±2.90
Average	88.92±2.81	84.34±2.95	97.80±3.21	62.22±7.61	91.58±1.90	95.54±2.81

that the cost of the generated solution is considerably less than that of the original dataset i.e. 1420 *PKR*. Furthermore, the accuracy of solution-1 as measured by F-measure has increased around 3 points as compared with the baseline results. Although the solution provided by test group selection is reasonably cost-effective, the feature set is comprised of a large number of features and the increase in the accuracy is not significant.

3) SOLUTION-2 (DIRECT FEATURE SELECTION)

Solution-2 is based on assigning cost values to each feature based on its relevance in a particular TG. In this regard, unlike solution-1, it is possible that a set of highly relevant features are selected and low relevance features are filtered out. As direct feature selection can be performed in two different ways i.e. *DFS-CT* and *DFS-TC*. Therefore, first, we report the results of *DFS-CT*. In this technique, 3 different

TABLE 13. *DFS-CT* detailed results (solution-2a).

Model	Accuracy	Precision	Recall	Specificity	F-measure	AUC
NB	95.70±4.30	93.80±6.30	100.0±0.0	88.60±11.10	94.70±3.30	96.40±5.10
LR	99.10±1.90	100.0±0.0	98.60±3.20	100.0±0.0	99.30±1.70	100.0±0.0
ANN	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
CART	90.40±5.60	92.60±7.80	93.0±8.80	85.80±16.0	92.30±4.4	97.10±3.30
RF	98.30±2.40	97.40±3.50	100.0±0.0	95.30±6.50	98.70±1.80	99.70±0.70
GBT	99.10±1.90	98.80±2.80	100.0±0.0	97.50±5.60	99.40±1.40	100.0±0.0
SVM	94.70±4.80	94.0±6.80	98.60±3.20	97.80±15.30	96.10±3.30	94.30±2.40
Average	96.75±2.98	96.65±3.88	98.60±2.17	93.57±7.78	97.63±2.27	98.21±1.64

TABLE 14. *DFS-TC* detailed results (solution-2b).

Model	Accuracy	Precision	Recall	Specificity	F-measure	AUC
NB	98.30±2.40	97.40±3.50	100.0±0.0	95.30±6.50	98.70±1.80	95.30±6.80
LR	89.60±7.90	92.40±8.10	91.50±9.30	85.60±15.90	91.60±6.40	97.10±3.20
ANN	99.10±1.90	100.0±0.0	98.60±3.20	100.0±0.0	99.30±1.70	100.0±0.0
CART	92.20±4.80	97.30±3.70	90.40±10.50	95.60±6.10	93.30±4.50	95.20±5.20
RF	98.30±2.40	97.30±3.70	100.0±0.0	95.60±6.10	98.60±1.90	100.0±0.0
GBT	95.60±3.10	96.20±5.60	97.20±3.80	93.10±10.10	96.60±2.40	77.80±2.58
SVM	96.50±1.90	97.40±3.50	97.20±3.80	95.30±6.50	97.20±1.60	99.0±2.10
Average	95.65±3.48	96.85±4.0	96.41±4.37	94.35±7.31	96.47±2.90	94.91±6.15

TABLE 15. Averaged results of individual feature scoring measures and their combination.

Method	Accuracy	Precision	Recall	Specificity	F-measure	AUC	Cost (*PKR*)
SU	93.52±4.61	92.10±5.98	98.80±1.74	84.81±12.37	95.10±3.42	98.88±1.54	1010
Chi-Sq.	93.64±2.62	92.60±3.11	93.21±3.12	92.64±5.71	93.21±2.15	0.94±2.28	1010
Relief	89.95±4.61	87.81±4.92	98.60±1.92	75.94±12.22	92.71±3.15	97.79±3.0	1010
Combined (*DFS-TC*)	95.65±3.48	96.85±4.0	96.41±4.37	94.35±7.31	96.47±2.90	94.91±6.15	1010

feature scoring lists are produced due to three feature scoring measures. The final list is obtained by taking the average value of each feature across different lists. A threshold value is selected based on the feature-cost curve as shown in Figure 7. As can be seen in Figure 7, the point is intersection around feature number 10. Therefore, the first 9 features are selected i.e. 15, 13, 19, 4, 20, 14, 6, 3, *and* 10. The overall cost of *DFS-CT* is 1010 *PKR*.

It can be seen in Table 14 that the overall accuracy of solution-2 (*DFS-CT*) is considerably more than that of solution-1 accuracy as provided in Table 13. Since solution-2 gives more preference to highly predicting features than the cost-effective features, therefore, the overall cost of solution-2 is greater than that of solution-1. Solution-2 can also be approached from a different direction i.e. applying threshold before combining the individual candidate solutions. In this, regard, *DFS-TC* provides a consolidated feature set solution based on individual feature subsets acquired from SU, Chi-Sq., and Relief. Table 14 provides details of the *DFS-TC* results, whereas the overall cost of the solution is 1010 PKR. The final feature set is comprised of 8 features i.e. 3, 4, 6, 13, 14, 15, 19, 20.

Solution-2 based on *DFS-TC* shows similar results as that of *DFS-CT*. Both these techniques incur the same cost i.e. 1010 PKR but there is a slight difference in their predictive accuracy values. Both the solutions are almost identical except *DFS-CT* includes an additional feature in the solution set i.e. feature number 10, blood glucose random, from TG 7. Since the cost of TG 7 is 0. Therefore, no extra cost is incurred for the *DFS-CT* solution. Both the aforementioned techniques show that highly predictive features are given more preference than cost-effective features as in the case of solution-1. Figure 10 depicts the feature-cost curve for the SU feature evaluation measure. The blue line represents feature score while the organ line depicts accumulated cost. The point of intersection is at feature number 3. Therefore, in Table 15 SU method evaluation is based on a solution having features 19, 20, 15, 13, 4, 14, 6, 22, and 3 as can be seen from the feature-cost graph is depicted in Figure 10.

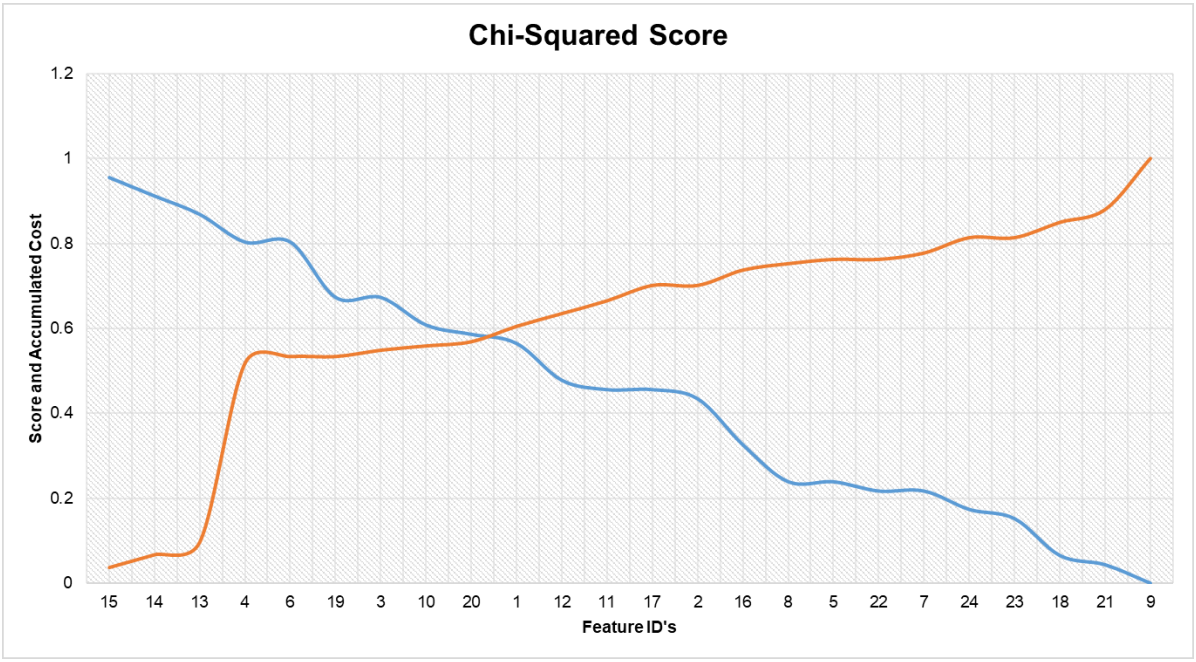

FIGURE 11. Feature-Cost curve for Chi-Square measure.

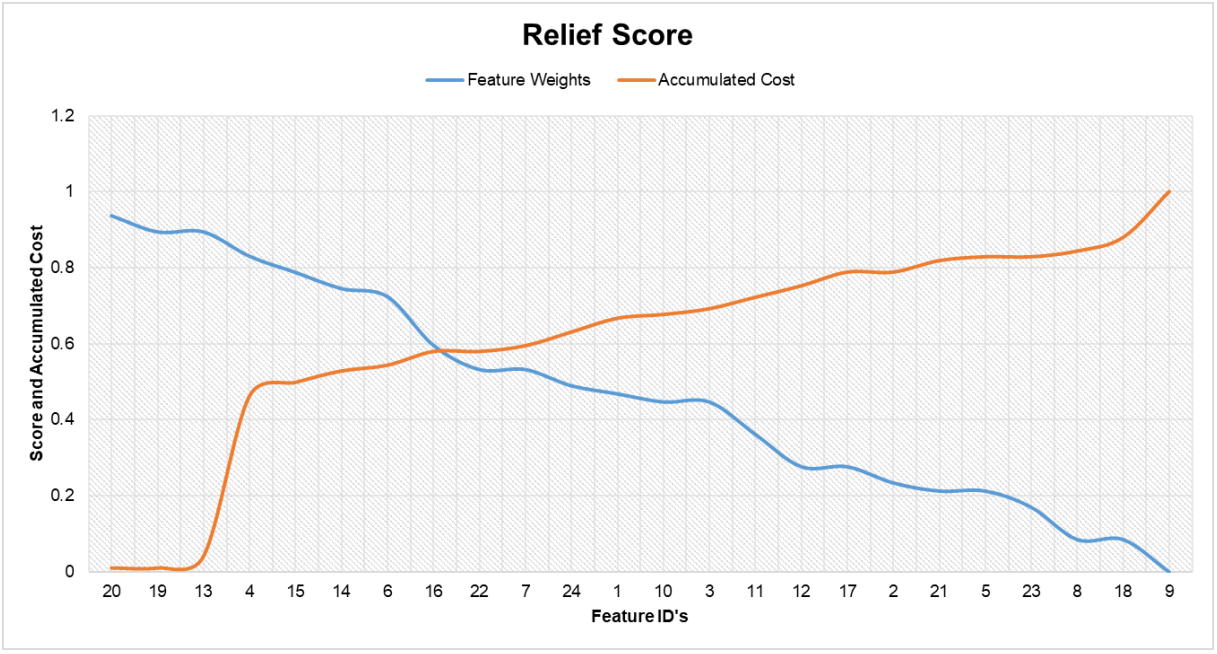

FIGURE 12. Feature-Cost curve for the Relief method.

Likewise, chi-square is used to assign feature weightage in Figure 11. In this case, the feature-cost curves intersect around feature number 20. Therefore, all the features above the threshold point are selected by this measure.

Finally, the Relief technique is used for feature scoring. In this case, the cut-off point of the threshold value is around feature number 16 as shown in Figure 12.

Furthermore, we report results for different configurations of *DFS-TC*. Table 15 provides averaged results over 7 classifiers for the threshold operation on the individual feature scoring measures and *DFS-TC* method that combines individual features subset using a multi-point intersection.

As can be seen in Table 15, the combined approach i.e. *DFS-TC* is slightly better than that of the individual measures on the CKD dataset. This observation is consistent with other

TABLE 16. Averaged results for set-theory based combination techniques for the *DFS-TC* approach.

Method	Accuracy	Precision	Recall	Specificity	F-measure	AUC	Cost (*PKR*)
Union	95.78±3.68	97.15±3.64	96.40±4.10	94.54±7.32	96.85±9.27	98.15±2.12	1010
Intersection	89.02±4.24	89.18±3.77	96.80±2.85	75.90±8.68	92.15±3.14	97.37±3.10	910
Multi-point Intersection	95.65±3.48	96.85±4.0	96.41±4.37	94.35±7.31	96.47±2.90	94.91±6.15	1010

TABLE 17. Combined feature set solution based on two candidate solutions.

Approach	No. of Features	Feature Set	Cost (*PKR*)	Average Accuracy	Run Time (Unit: s)
Solution-1	14	3, 6, 7, 8, 15, 16, 17, 18, 24, 13, 14, 22, 9, 23	460	88.92±2.81	3.04
Solution-2	8	3, 4, 6, 13, 14, 15, 19, 20	1010	96.75±2.98	2.48
Combined	5	3, 6, 13, 14, 15	260	91.25±3.41	3.19

TABLE 18. Comparison between proposed and other related approaches for cost-sensitive ensemble feature selection.

Approach	Accuracy	Precision	Recall	Specificity	F-measure	AUC	Features	Cost (*PKR*)
[29]	90.0±2.80	90.71±2.57	97.22±2.60	78.25±4.31	93.07±2.22	97.94±2.12	9	1210
[30]	87.37±3.40	86.84±3.47	98.62±1.50	68.87±9.18	92.95±2.42	97.60±3.01	11	1110
[22]	91.74±2.32	92.55±3.02	97.58±2.87	82.15±4.72	94.34±1.95	99.01±0.15	5	1010
[13]	85.74±2.44	84.92±2.91	99.40±1.0	62.94±4.84	90.75±1.82	91.84±2.98	11	1110
[21]	93.41±3.27	94.24±3.04	96.61±3.34	88.20±6.21	95.0±2.64	98.11±2.55	12	1010
[10]	83.29±5.07	82.66±5.82	97.20±3.34	60.04±12.11	88.56±3.75	93.40±5.80	13	910
[12]	92.64±2.74	93.50±2.45	97.39±3.02	84.69±5.44	94.41±2.15	98.70±1.50	8	1010
[28]	87.37±3.40	86.84±3.47	98.62±1.50	68.87±9.18	92.95±2.42	97.60±3.01	11	1110
Solution-1	88.92±2.81	84.34±2.95	97.80±3.21	62.22±7.61	91.58±1.90	95.54±2.81	14	460
Solution-2	96.75±2.98	96.65±3.88	98.60±2.17	93.57±7.78	97.63±2.27	98.21±1.64	8	1010
Combined	91.25±3.41	97.81±5.28	91.98±56.12	97.37±87.14	94.81±2.81	96.55±2.37	5	260

ensemble studies where the overall solution is generally better than the solutions provided by the ensemble's elements. Furthermore, it is important to note that the combination method employed in *DFS-TC* is based on a multi-point intersection. In Table 16 reports averaged results over 7 classifiers for different combination methods.

It can be seen that different set-combining techniques have different results. In the case of union, the overall accuracy is high but the size of the final solution is 11. In the case of an intersection, the overall cost is low with 7 features in its solution set but the accuracy is also decreased. The multi-point intersection provides a trade-off where the cost and accuracy are almost the same as that of the union while the solution set is comprised of 8 features. Therefore, the multi-point intersection technique is favored for the *DFS-TC* approach. As *DFS-CT* is selected for solution-2, therefore, both solution-1 and solution-2 feature sets can be combined into a single solution as shown in solution schemata in Figure 1. The candidate feature subsets are obtained from both of the solutions that can be combined using intersection operation. Table 17 shows features in solution-1, solution-2, and the combined solution. The operations detailed in Algorithm 1 $\sim$ 2 are of two types i.e. standard operations such as sorting, selection, merging, etc., and operations related to executing the FSF that is composed of ensemble measures. Although the FSF is the most expensive operation, it can be performed in a parallel fashion on multiple data partitions. Furthermore, the final solution (combined method) relies on solution-1 and solution-2 that can also be executed in parallel. The averaged run-time of the 5 executions for each solution-1, solution-2 and the final solution (combined) is reported in Table 17.

The proposed approach is compared with other feature selection methods reported in the literature. Some of these methods have reported results on the CKD dataset while other methods perform similar operations following the ensemble feature selection methodology. Table 18 provides details of the comparison.

Table 18 report results for proposed methods and other relevant feature selection methods. In this regard, as it can be seen that solution-1 provided a feature subset with the lowest cost while its accuracy is comparable to some of the feature selection methods. In terms of feature set cardinality both combined solution and Salekin and Stankovic [22] yield compact feature sets, while solution-2 produced the largest feature set among all the methods compared in this research. Solution-2 generated a highly accurate feature subset having cost comparable to that of other methods. The combined approach is a trade-off solution having comparably low accuracy than solution-2 but the overall cost is

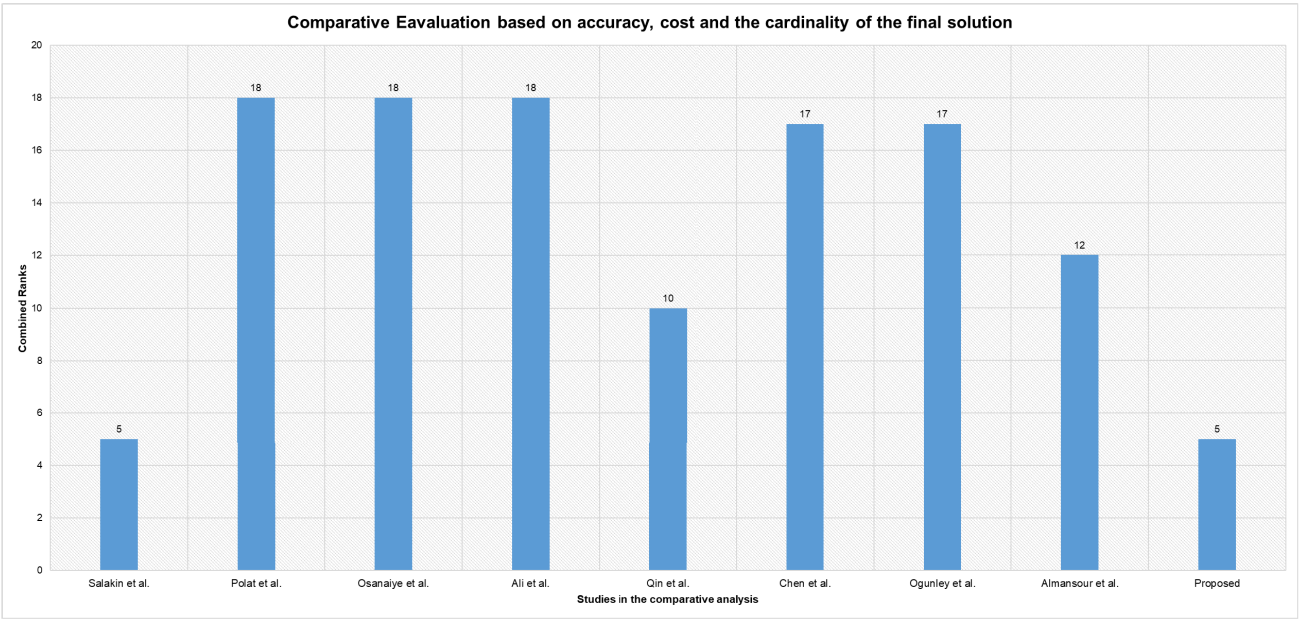

FIGURE 13. Comprehensive evaluation of the proposed approach with that of different comparative methods.

TABLE 19. Statistical evaluation of the comparative methods.

Approach	Statistical Difference
Chen *et al.,* [29]	∗
Ali *et al.,* [30]	∗
Salekin *et al.,* [22]	∗
Polat *et al.,* [13]	∗
Almansour *et al.,* [21]	∗
Ogunley *et al.,* [10]	†
Qin *et al.,* [12]	∗
Osanaiye *et al.,* [28]	∗
∗ Proposed solution is not statistically better	
† Proposed solution is statistically better	

also low. It is important to note that all the results reported in Table 18 are based on the CKD diagnosis dataset. The difference between the proposed method and other feature selection methods in terms of accuracy is not significant, except XGBoost, as reported in Table 19. Therefore, most of the techniques provide sufficiently accurate results on the CKD dataset while it can also be seen that the overall cost of a solution varies from 260 PKR to 1210 PKR. Therefore, cost emerges as an important consideration that differentiates the proposed approach based on cost-sensitive feature selection from that of other feature selection methods.

Finally, a comparative analysis is depicted in Figure 13. The comparison is based on three characteristics i.e. F-measure, cost of the selected feature set, and the cardinality of the final feature set solution. In this regard, lower rank shows better performance. It is pertinent to note that both Salekin and Stankovic [22] and the proposed approach obtain the same rank in terms of the aforementioned ranking criteria. Moreover, among all the other feature selection techniques, both [22] and the proposed approach explicitly cater to the cost-factor in the final feature set selection. Although the rank of the proposed approach is the same as that of [22], the individual difference in terms of incurred cost is more pronounced.

V. DISCUSSION

In this section, we discuss the research questions raised in Section I in the light of the experimentation and results reported in Section IV. In this regard, we will draw conclusions based on the empirical evaluation in an attempt to address the pressing research questions. It is pertinent to note that all this discussion about material and methods is within the scope of chronic kidney disease dataset. The main objective of this research as mentioned in Section I is to investigate the practical aspects of the machine learning-based automated decision support systems. In this regard, it is observed that a vast number of studies are geared towards enhancing the predictive accuracy of the decision models with little regard to the overall cost of operating the proposed solutions. This study focuses on an often overlooked yet of a high significance area of cost-sensitive feature selection. In this study, we are concerned with acquiring the actual cost of data acquisition rather than simulating the cost factor in order to draw meaningful conclusions. In this regard, we have opted for CKD as a case study and developed this study from the developing countries' perspective. It is also important to note that the cost factor associated with candidate solutions

is in terms of conducting medical tests for acquiring measurements or data for the required features mentioned in the CKD dataset [34]. In this section the following questions are discussed in the light of the extensive experimentation performed for the CKD diagnosis problem:

A. RESEARCH QUESTIONS

A. How to assign a proxy relevance score to a TG and a proxy cost value to a feature? The relevance score cannot be directly assigned to a test group i.e. in the case of solution-1. Therefore, each test group can be assigned a proxy relevance based on features' weight-age. In this study, three filter methods are used for assigning a feature score that represents a feature's collective worth in the feature set. A test group can be assigned a proxy relevance score as a ratio of the overall score of its member features to the overall score of the feature set. As the problem formulation deals with a group of features rather than individual features, therefore, the cost is directly associated with a particular test group. We have demonstrated two approaches for proxy cost assignment to a feature. The first approach directly deals with test group selection, as reported in Table 10. While the second approach deals with assigning a cost factor based on a feature's contribution in the prediction step. In this regard, the cost is assigned relative to a feature's weight-age in a given test group i.e. higher the feature weight-age, the lower the proxy cost value (refer to Table 13 and Table 14). Both of these approaches impact the feature selection process. The first approach, solution-1, of cost assignment at the test group level leads to a relatively lower cost solution than the solution-2. In this regard, the final feature subset obtained by solution-1 is comprised of a large feature set than that of solution-2 because once a test group is selected then all the features that are members of the selected test group are also implicitly selected as well. This selection mechanism may result in model over-fitting and eventually, the overall model accuracy on test data may also be affected.

B. How to select a threshold value for feature subset selection? Filter based feature selection produces an ordered list of features with respect to their relevance. In this regard, the application of a threshold selection becomes necessary, unlike subset selection methods. As it is mentioned in Section II that different threshold selection techniques are available i.e. fixed threshold value and threshold selection based on data complexity measures. In this study, two aspects of feature selection are taken into consideration i.e. weight-age of a feature and its cost. In this regard, a heuristic is employed that uses a feature-cost curve to find a point of intersection. This point of intersection serves as a potential candidate for a threshold value i.e. features beyond the threshold point are not deemed as cost-effective. In order to find a threshold point both feature weight-age and the accumulated cost values are normalized as shown in Figure 6 and Figure 7.

C. How to combine candidate solutions into a consolidated solution? Solution-2 deals with two techniques i.e. *DFS-CT* and *DFS-TC*. The *DFS-TC* deals with applying threshold operation to individual feature scoring measures that essentially results in a set of candidate feature subsets. As this study employed 3 feature scoring techniques, therefore, the resultant candidate subsets are also three. Once candidate feature subsets are acquired, the next important aspect is combining these individual solutions into a consolidated solution. In this regard, multiple set operations are at our disposal such as intersection, union, and, multi-point intersection. We demonstrated in Section IV that intersection results in a compact solution and lower cost at the expense of predictive accuracy. It is due to the reason that if a feature does not make it to any of the candidate solutions, then it is discarded. Some of the features may provide complementary information in interaction with other features and on their own, they are not very informative. Union approach resulted in a larger solution set because it gathers all the features present in candidate solutions. In the worst-case scenario, the union approach may result in almost all the features present in the original dataset, it is due to the fact that the union approach does not employ any conflict resolution strategy. In the case of a multi-point intersection that acts as a majority-voting operation, a trade-off solution is acquired i.e. it has a relatively compact solution set while the overall accuracy and cost are comparable to that of the union technique (refer to Table 16).

B. COMPARATIVE ANALYSIS

A comparative result of the proposed method with that of other related methods is provided in Table 18. It is pertinent to note that the reported results are averaged over multiple classifiers, mentioned in Table 9.There are two important considerations in this regard i.e. accuracy of the solution and cost of the solution. Ideally, the final solution should have the highest accuracy and the lowest cost. As it is consistent with the observation that the high-cost features tend to have high accuracy as well, therefore, a trade-off solution is to be selected with reasonably high accuracy and a low cost. In this study, as we are dealing with the cost that is associated with a test group rather than directly with the feature, therefore, the solution needs to select features that have an overall lower test group cost while at the same time demonstrating good discriminating power that is subsequently levered by the classification algorithm. The combined method is the final proposed solution while both solution-1 and solution-2 serve as intermediate solutions. In terms of F-measure the solution-2 achieved the highest accuracy while also incurring cost comparable with that of other methods. Both solution-1 and combined method produced feature sets with lower overall cost compared to the remaining methods presented in this study. In terms of AUC, Salekin and Stankovic [22] achieved the highest score of 99.01 ± 0.15 while only selecting 5 features. The proposed combined method also selects 5 features over the CKD dataset while lowering the overall incurred cost to 260 PKR. This result can be explained through the design approach of the proposed combined method that select features in a group-aware manner while the [22] selects feature directly without any consideration to the feature's membership to a group.

C. MULTI-PRONGED APPROACH

Furthermore, in this study, a cost-sensitive ensemble feature selection method is proposed for group-based features. It is demonstrated that a straight-away manner of test group selection can produce a cost-effective solution but it suffers from two issues i.e. low predictive accuracy and high cardinality of the final solution (refer to Table 12). Afterward, another approach of cost-sensitive feature selection is demonstrated that can successfully select highly predictive features at the expense of a higher solution cost (refer to Table 13). Two different techniques are demonstrated for solution-2 differentiated on the application of the threshold operation. Based on the empirical evaluation (refer to Table 13 and Table 14) both these techniques produce almost similar results based on predictive accuracy, cost, and cardinality of the feature subset. A combined solution is devised based on solution-1 and solution-2 (refer to Table 17). The combined solution retains some of the characteristics of solution-1 e.g. low cost and similarly, it also reflects characteristics of solution-2 e.g. high predictive accuracy. It is important to note that the main objective of this study is to select a cost-effective solution for the CKD diagnosis problem. In this regard, the accuracy of the proposed methods as compared with other feature selection methods is not statistically significantly better (refer to Table 19) as the cost-independent feature selection methods are supposed to select highly predictive features without any regard to the incurred cost. It is demonstrated that comparable accuracy over the CKD dataset can be obtained while also catering to the cost aspect of the solution as well. Moreover, both the solution-1 and the combined method result in one of the lowest cost solutions among the other comparable methods. Figure 13 depicts the results of comparative techniques with the final combined solution. It can be seen that both Salekin and Stankovic [22] and the proposed combined solution has secured the same rank. Both the aforementioned techniques deal with cost-sensitive feature selection for the CKD dataset. Although the ranks of both these techniques are the same in terms of cost value, the proposed combined approach is better than [22]. In this regard, from a developing countries perspective, it is paramount to decrease the overall operating cost of automated diagnosis systems. Moreover, these systems serve as the first line of defense by providing the necessary patient screening capabilities therefore the cost of application, and its intended benefit is of special consideration by the decision-makers e.g. hospital administration.

D. SCALABILITY PERSPECTIVE

The proposed ensemble approach is primarily composed of univariate and multivariate filter techniques. In this regard, it is extensively reported in the literature that filter methods are generally more scalable than their counterpart wrapper methods [31], [42], [43].We have alluded to this observation in Table 3. Due to this reason, in the biomedical domain where the microarray gene classification datasets generally range from 1000 $\sim$ 10,000 genes, the filter methods are preferred due to their scalability to very high-dimensional datasets, computational simplicity, and less computational complexity. It is reported that on very high-dimensional datasets, the wrapper methods employing classification algorithms as a candidate solution evaluation function tend to degrade when faced with a high number of irrelevant features [31]. Furthermore, this research deals with the selection of features that are highly predictive as well as cost-effective. Therefore, we have selected the benchmarked chronic kidney disease dataset and acquired the respective cost factor from a developing country. This dataset does not pose the challenge of scalability, therefore, both filter and wrapper methods perform equally well on this dataset as reported in Table 18. The filter approach provides a holistic picture where all the features are ranked according to their predictive scores. The cost factor is leveraged to select an appropriate threshold value. In this regard, the wrapper methods only select a subset of features while the rest of the features are discarded. Therefore, the relative importance of the features is not available to the decision-maker to externally validate the results from the domain knowledge. In light of the above discussion, we can conclude that as filter methods are generally more scalable than the wrapper methods and the proposed ensemble approach is based on filter methods, therefore, it is relatively more scalable to a similar ensemble approach using the wrapper-based feature selection techniques.

E. COMPUTATIONAL PERSPECTIVE

We have alluded to the general computational aspects of filter methods in Table 18. Generally, filter methods are less computationally expensive than the wrapper methods, therefore, inherently the ensemble approach based on filter techniques would also be less expensive than that of wrapper methods in an ensemble configuration. Moreover, as the ensemble components can be executed in an independent manner, therefore, the overall cost of the feature scoring function (FSF) would be as per the slowest component of the ensemble. Furthermore, the time required to calculate the automatic threshold is almost negligible, especially for datasets with small to medium sized datasets. The operations to join lists and subsets are based on averaging and majority-voting (multi-point interaction) computations, respectively, therefore, the unnecessary computational overheads are also avoided.

F. DESIGN CHOICES

The proposed filter-based ensemble approach employs two univariate (Symmetric Uncertainty, Chi-Squared) and one bi-variate (Relief) technique as a base feature scoring measure. In this regard, the parameter-sensitivity is different than that of the population-based wrapper approaches e.g. genetic algorithm, particle swarm optimization, ant colony optimization, etc., where an extensive strategy is required for the parameter selection e.g. the population size, reproduction operators, number of generations, among others. The proposed approach leverages statistical and information-theoretic measures for quantifying the relevance of features in an ensemble configuration. The default parameters are used

for the operations such as the size of the neighborhood and the number of bins i.e. 10, required by the aforementioned measures. Furthermore, the accumulated cost is computed using the meta-data information provided along with the dataset and does not require any external tunable parameters. In Table 8, we have provided the default parameters used in constructing the classification models. Please note that no parameter tuning is performed for the classification models in order to report a fair comparison between the proposed and the other approaches mentioned in the manuscript. In this regard, the baseline models are used with the default parameters provided by the RapidMiner data science simulation software. There are two design decisions taken in the proposed approach:

- Application of the threshold operation in the solution-2 - Direct Feature Selection (DFS). In the proposed methodology, the threshold can be applied after combining the individual solutions (DFS-CT) and before combining the individual solutions (DFS-TC). We have reported results of both these options in Table 13 and Table 14. Based on the empirical results, the DFS-CT technique is selected, although the difference between both the aforementioned techniques is not significant.
- Combining multiple feature subsets into a consolidated feature set. In case the DFS-TC is selected, then there are multiple options to obtain a final solution e.g. union, intersection, and multi-point intersection. In Table 16 we have reported results for each set combining technique along with the incurred cost of the final solution. Although both multi-point intersection and union result in solutions having comparable accuracy, the former technique tends to select a feature set with lower cardinality as compared with the latter technique. This empirical observation is supported by several works where the majority vote option is preferred over others [10], [27], [29].

G. STABILITY PERSPECTIVE

Chronic kidney disease dataset under consideration did have certain deficiencies in terms of missing values. Apart from it, the dataset contained a few outliers. As a number of comparative studies using the CKD dataset only relied on data imputation, therefore, the dataset is not treated for any other elaborate preprocessing operations. Although the CKD dataset provided reasonably high accuracy with limited data preprocessing, the proposed methodology uses an effective multi-step feature scoring technique to deal with spurious patterns as illustrated in Figure 5. In order to deal with noise in the data, multiple horizontal data partitions are created. In this study, the number of partitions is set according to the number of classes present in the dataset. A set of feature scoring measures is applied to each data partition and a subsequent list containing the feature ranking is obtained. As the proposed approach is based on the heterogeneous ensemble method, therefore, the feature scoring technique is aimed at addressing the issue of noise due to spurious patterns and inducing diversity in the sets of selected features, thereby improving performance and obtaining more robust and stable solutions. It is important to note that the proposed approach requires cost information associated with the feature groups where features are distributed into non-overlapping groups.

VI. CONCLUSION

In this study, we have focused on an important yet generally overlooked area of machine learning applications. In this regard, we have focused on the cost of operating a machine learning solution within the scope of developing countries. Therefore, the case study problem is based on a benchmark dataset that is used in several recent studies, and the respective cost factor is also taken from a developing world perspective. Furthermore, unlike some of the studies on cost-based feature selection where an overly simplified version of cost is considered i.e. interdependence of features are overlooked, we have modeled the problem as a non-overlapping group-based feature selection. Due to the aforementioned problem formulation, two approaches are devised to address issues of incurred cost and predictive accuracy, respectively. Through extensive experimentation on the benchmarked CKD dataset, it is demonstrated that both the proposed approaches for group-based feature selection can be successfully combined into a consolidated final solution. The stated combined solution is, thereafter, compared with 8 feature selection methods over a set of 7 classifiers. Based on the predictive accuracy results it is observed that in general, the proposed approach is not statistically significantly better than other methods. Although some of the comparative methods have higher predictive accuracy than the proposed method on the CKD dataset, in general, this is consistent with the observation that other methods are designed to maximize the accuracy metric while the proposed method is slanted towards a trade-off solution. The overall cost of the proposed method is decreased by a factor of 5.57 i.e. the cost of the final solution is 260 PKR from 1420 PKR, while the total number of features is reduced by a factor of 4.80 i.e. selected features are 5 out of 24. In this regard, it is demonstrated that the overall cost incurred by the proposed solution is considerably lower than that of the other comparative techniques while the predictive accuracy remained reasonably high.

This study may be extended in a number of directions such as the employed ensemble feature ranking approach can be enhanced to address the feature-feature interaction problem, as the feature dependency is not adequately captured by filter techniques.

The benchmarked CKD dataset used in this study is relatively noise-free. Although the opportunity did not present itself to evaluate the impact of noisy data in the CKD dataset, we believe it would be an interesting research direction to investigate the efficacy of the proposed approach on the high-dimensional noisy data that also contains associated feature-group cost information.

Furthermore, cost-sensitive feature selection problem may be re-stated as a multi-objective problem where a set of trade-

off solutions are provided to the decision-maker in order to provide a wider view of the solution space. Finally, yet another fruitful direction to extend this study is to consider, apart from economic cost, other cost factors such as availability of a resource, computational time required to acquire data, nature of risk associated with data acquisition, among others.

REFERENCES

[1] A. S. Levey, J. Coresh, K. Bolton, B. Culleton, K. S. Harvey, T. A. Ikizler, C. A. Johnson, A. Kausz, P. L. Kimmel, and J. Kusek, "K/DOQI clinical practice guidelines for chronic kidney disease: Evaluation, classification, and stratification," *Amer. J. Kidney Diseases*, vol. 39, no. 2, pp. S1–266, 2002.

[2] J. A. Vassalotti, L. A. Stevens, and A. S. Levey, "Testing for chronic kidney disease: A position statement from the national kidney foundation," *Amer. J. Kidney Diseases*, vol. 50, no. 2, pp. 169–180, Aug. 2007.

[3] J. Borawski, M. Wilczynska-Borawska, W. Stokowska, and M. Mysliwiec, "The periodontal status of pre-dialysis chronic kidney disease and maintenance dialysis patients," *Nephrol. Dialysis Transplantation*, vol. 22, no. 2, pp. 457–464, Sep. 2006.

[4] M. Hasan, I. Sutradhar, R. D. Gupta, and M. Sarker, "Prevalence of chronic kidney disease in South Asia: A systematic review," *BMC Nephrol.*, vol. 19, no. 1, p. 291, 2018.

[5] S. Varughese and G. Abraham, "Chronic kidney disease in India: A Clarion call for change," *Clin. J. Amer. Soc. Nephrol.*, vol. 13, no. 5, pp. 802–804, 2018.

[6] S. Imtiaz, B. Salman, R. Qureshi, M. Drohlia, and A. Ahmad, "A review of the epidemiology of chronic kidney disease in pakistan: A global and regional perspective," *Saudi J. Kidney Diseases Transplantation*, vol. 29, no. 6, p. 1441, 2018.

[7] R. A. Nugent, S. F. Fathima, A. B. Feigl, and D. Chyung, "The burden of chronic kidney disease on developing nations: A 21st century challenge in global health," *Nephron Clin. Pract.*, vol. 118, no. 3, pp. c269–c277, 2011.

[8] S. Imran Ali, B. Ali, J. Hussain, M. Hussain, F. A. Satti, G. H. Park, and S. Lee, "Cost-sensitive ensemble feature ranking and automatic threshold selection for chronic kidney disease diagnosis," *Appl. Sci.*, vol. 10, no. 16, p. 5663, Aug. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/16/5663

[9] A. Sobrinho, A. C. M. D. S. Queiroz, L. Dias Da Silva, E. De Barros Costa, M. Eliete Pinheiro, and A. Perkusich, "Computer-aided diagnosis of chronic kidney disease in developing countries: A comparative analysis of machine learning techniques," *IEEE Access*, vol. 8, pp. 25407–25419, 2020.

[10] A. A. Ogunleye and W. Qing-Guo, "XGBoost model for chronic kidney disease diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Apr. 17, 2019, doi: 10.1109/TCBB.2019.2911071.

[11] M. S. Wibawa, I. M. D. Maysanjaya, and I. M. A. W. Putra, "Boosted classifier and features selection for enhancing chronic kidney disease diagnose," in *Proc. 5th Int. Conf. Cyber IT Service Manage. (CITSM)*, Aug. 2017, pp. 1–6.

[12] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2020.

[13] H. Polat, H. Danaei Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," *J. Med. Syst.*, vol. 41, no. 4, p. 55, Apr. 2017.

[14] A. A. Freitas, "Comprehensible classification models: A position paper," *ACM SIGKDD Explor. Newslett.*, vol. 15, no. 1, pp. 1–10, 2014.

[15] S. Itani, F. Lecron, and P. Fortemps, "Specifics of medical data mining for diagnosis aid: A survey," *Expert Syst. Appl.*, vol. 118, pp. 300–314, Mar. 2019.

[16] Q. Zhou, H. Zhou, and T. Li, "Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features," *Knowl.-Based Syst.*, vol. 95, pp. 1–11, Mar. 2016.

[17] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Comput. Biol. Med.*, vol. 112, Sep. 2019, Art. no. 103375.

[18] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Inf. Sci.*, vol. 484, pp. 153–166, May 2019.

[19] X. Song, L. R. Waitman, Y. Hu, A. S. L. Yu, D. Robins, and M. Liu, "Robust clinical marker identification for diabetic kidney disease with ensemble feature selection," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 3, pp. 242–253, Mar. 2019.

[20] B. Pes, "Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains," *Neural Comput. Appl.*, vol. 32, no. 10, pp. 5951–5973, May 2020.

[21] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, S. Alrashed, and S. O. Olatunji, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Comput. Biol. Med.*, vol. 109, pp. 101–111, Jun. 2019.

[22] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Oct. 2016, pp. 262–270.

[23] X. Gu, Y. Li, and J. Jia, "Feature selection for transient stability assessment based on kernelized fuzzy rough sets and memetic algorithm," *Int. J. Electr. Power Energy Syst.*, vol. 64, pp. 664–670, Jan. 2015.

[24] Y. Li and Z. Yang, "Application of EOS-ELM with binary jaya-based feature selection to real-time transient stability assessment using PMU data," *IEEE Access*, vol. 5, pp. 23092–23101, 2017.

[25] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egyptian Informat. J.*, vol. 19, no. 3, pp. 179–189, Nov. 2018.

[26] B. Seijo-Pardo, V. Bolón-Canedo, A. Alonso-Betanzos, "Using data complexity measures for thresholding in feature selection rankers," in *Advances in Artificial Intelligence. CAEPIA* (Lecture Notes in Computer Science), vol. 9868, O. Luaces *et al.*, Eds. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-44636-3_12.

[27] B. Seijo-Pardo, V. Bolón-Canedo, and A. Alonso-Betanzos, "On developing an automatic threshold applied to feature selection ensembles," *Inf. Fusion*, vol. 45, pp. 227–245, Jan. 2019.

[28] O. Osanaiye, H. Cai, K.-K. R. Choo, A. Dehghantanha, Z. Xu, and M. Dlodlo, "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, p. 130, 2016.

[29] C.-W. Chen, Y.-H. Tsai, F.-R. Chang, and W.-C. Lin, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results," *Expert Syst.*, vol. 37, no. 5, 2020, Art. no. e12553.

[30] M. Ali, S. I. Ali, D. Kim, T. Hur, J. Bang, S. Lee, B. H. Kang, and M. Hussain, "uEFS: An efficient and comprehensive ensemble-based feature selection methodology to select informative features," *PLoS ONE*, vol. 13, no. 8, 2018, Art. no. e0202705.

[31] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Inf. Fusion*, vol. 52, pp. 1–12, Dec. 2019.

[32] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Distributed feature selection: An application to microarray data classification," *Appl. Soft Comput.*, vol. 30, pp. 136–150, May 2015.

[33] S. I. Ali and S. Lee, "Ensemble based cost-sensitive feature selection for consolidated knowledge base creation," in *Proc. 14th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2020, pp. 1–7.

[34] D. Dua and C. Graff, "UCI machine learning repository," Ph.D. dissertation, School Inf. Comput. Sci., Univ. California, Irvine, CA, USA, 2019.

[35] S. Wen, W. Liu, Y. Yang, P. Zhou, Z. Guo, Z. Yan, Y. Chen, and T. Huang, "Multilabel image classification via Feature/Label co-projection," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Feb. 6, 2020, doi: 10.1109/TSMC.2020.2967071.

[36] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, and Z. Zeng, "CLU-CNNs: Object detection for medical images," *Neurocomputing*, vol. 350, pp. 53–59, Jul. 2019.

[37] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.

[38] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Informat.*, vol. 85, pp. 189–203, Sep. 2018.

[39] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*, vol. 207. Berlin, Germany: Springer-Verlag, 2008.

[40] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "Yale: Rapid prototyping for complex data mining tasks," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2006, pp. 935–940.

[41] M. G. Kendall, "Rank correlation methods," Tech. Rep., 1948.

[42] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.
[43] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "An ensemble of filters and classifiers for microarray data classification," *Pattern Recognit.*, vol. 45, no. 1, pp. 531–539, Jan. 2012.

SYED IMRAN ALI received the B.S. degree in computer science from IQRA University, Islamabad, in 2008, and the M.S. degree in computer science from the National University of Computer and Emerging Sciences (NUCES), in 2012. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Kyung Hee University, South Korea. He has taught at NUCES as an Adjunct Faculty Member. His research interests include machine learning, health analytics, and data-driven systems.

HAFIZ SYED MUHAMMAD BILAL received the M.S. degree in computer science from the National University of Sciences and Technology, Pakistan, in 2008. He is currently pursuing the Ph.D. degree in computer science and engineering with Kyung Hee University, South Korea. He has working experience of more than three years in data science and open-source development and is actively involved in developing big data ecosystem for academic and health care. His research interests include behavior quantification and assessment, machine learning, and behavior modeling and adaptation.

MUSARRAT HUSSAIN received the M.S. degree in software engineering from the National University of Science and Technology (NUST), Pakistan, in 2015. He is currently pursuing the Ph.D. degree with the Ubiquitous Computing Laboratory, Kyung Hee University, South Korea. His research interests include clinical text mining, knowledge extraction and representation, and machine learning.

JAMIL HUSSAIN received the Ph.D. degree from the Department of Computer Engineering, Kyung Hee University, South Korea, in 2019. He is currently working as a Postdoctoral Researcher with the Ubiquitous Computing Laboratory, Kyung Hee University. He has a professional experience of over seven years in software industry working on user experience design and development on various projects. His research interest includes user experience design, artificial intelligence, and information extraction from textual data.

FAHAD AHMED SATTI received the M.S. degree in computer science from the University of Trento, Italy, in 2014. He is currently pursuing the Ph.D. degree in computer engineering with Kyung Hee University, South Korea. His primary research interest is directed toward providing a technical solution to resolve the lack of interoperability in information systems. In general, he is interested in domains of semantic matching, stream reasoning, knowledge extraction, and machine learning.

MAQBOOL HUSSAIN received the B.S. degree in computer science from the Kohat University of Science and Technology, Kohat, Pakistan, the M.S. degree from the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Pakistan, in 2009, and the Ph.D. degree from Kyung Hee University, South Korea, in 2016. He is currently an Assistant Professor with Sejong University, South Korea. He is also working as a Visiting Scholar with Oakland University, MI, USA. He has more than eight years of software development experience and is involved in consultation and development of hospital system with collaboration of medical experts. His research interests include artificial intelligence, and healthcare interoperability standards and clinical decision support systems.

GWANG HOON PARK (Senior Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Yonsei University, Seoul, South Korea, in 1985 and 1987, respectively, and the M.S. and Ph.D. degrees in electrical engineering and applied physics from Case Western Reserve University, OH, USA, in 1991 and 1995, respectively. He was a Principal Research Engineer with the Information and Telecommunication Research and Development Center, Hyundai Electronics Industries, Icheon, South Korea, from 1995 to 1997, and an Associate Professor with the Department of Computer Science, Yonsei University, Wonju, South Korea, from 1997 to 2001. Since 2001, he has been a Professor with the Department of Computer Engineering, Kyung Hee University, South Korea. His research interests include video signal processing, multimedia systems, pattern recognition, and computational intelligence.

TAECHOONG CHUNG received the B.S. degree in electronic engineering from Seoul National University, Republic of Korea, in 1980, and the M.S. and Ph.D. degrees in computer science from KAIST, Republic of Korea, in 1982 and 1987, respectively. Since 1988, he has been with the Department of Computer Science and Engineering, Kyung Hee University, Republic of Korea, where he is currently a Professor. His research interests include machine learning, meta search, and robotics.

SUNGYOUNG LEE (Member, IEEE) received the B.S. degree from Korea University, Seoul, South Korea, and the M.S. and Ph.D. degrees in computer science from the Illinois Institute of Technology, Chicago, IL, USA, in 1987 and 1991, respectively. He was an Assistant Professor with the Department of Computer Science, Governors State University, University Park, IL, USA, from 1992 to 1993. Since 1993, he has been a Professor with the Department of Computer Engineering, Kyung Hee University, South Korea, where he has been the Director of the Neo Medical ubiquitous-Life Care Information Technology Research Center, since 2006. He is currently the Founding Director of the Ubiquitous Computing Laboratory. His current research interests include ubiquitous computing and applications, wireless ad hoc and sensor networks, context-aware middle ware, sensor operating systems, real-time systems and embedded systems, and activity and emotion recognition. He is a member of ACM.