Received: 17 January 2022 | Revised: 4 May 2022 | Accepted: 23 May 2022

DOI: 10.4218/etrij.2022-0018

ORIGINAL ARTICLE

ETRI Journal WILEY

Relevancy contemplation in medical data analytics and ranking of feature selection algorithms

P. Antony Seba[1] | **J. V. Bibal Benifa**[2]

[1]Department of Computer Science and Engineering, Indian Institute of Information Technology Kottayam, Kottayam, Kerala, India

[2]Department of Computer Science and Engineering, Information Technology Kottayam, Kottayam, Kerala, India

Correspondence
P. Antony Seba, Department of Computer Science and Engineering, Indian Institute of Information Technology Kottayam, Kottayam, Kerala, India.
Email: sebaantony.phd201002@iiitkottayam.ac.in

Funding information
This research was not supported by any funding.

Abstract

This article performs a detailed data scrutiny on a chronic kidney disease (CKD) dataset to select efficient instances and relevant features. Data relevancy is investigated using feature extraction, hybrid outlier detection, and handling of missing values. Data instances that do not influence the target are removed using data envelopment analysis to enable reduction of rows. Column reduction is achieved by ranking the attributes through feature selection methodologies, namely, extra-trees classifier, recursive feature elimination, chi-squared test, analysis of variance, and mutual information. These methodologies are ranked via Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) using weight optimization to identify the optimal features for model building from the CKD dataset to facilitate better prediction while diagnosing the severity of the disease. An efficient hybrid ensemble and novel similarity-based classifiers are built using the pruned dataset, and the results are thereafter compared with random forest, AdaBoost, naive Bayes, k-nearest neighbors, and support vector machines. The hybrid ensemble classifier yields a better prediction accuracy of 98.31% for the features selected by extra tree classifier (ETC), which is ranked as the best by TOPSIS.

KEYWORDS
data contemplation, DEA, feature selection, TOPSIS

1 | INTRODUCTION

Machine learning models are developed to make accurate predictions, and such outcomes are highly mandated in medical data analytics. Such models are expected to exploit all the instances and features of a dataset to facilitate an unambiguous contribution toward making appropriate decisions [1]. Data have been identified as the most important part of machine learning, and the core concept of data science investigates data purity to yield better predictions. Data represent certain characteristics, and these can be used to extract meaningful insights to provide feasible solutions for real-world problem through analytics. The quality of data samples makes a significant contribution to a successful training process and improves the performance of the predictive models. Data contemplation utilizes scientific methods to extract relevant structured data using advanced data analysis for decision-making.

In general, the relevancy scrutiny of data instances and attributes has been seen to improve the process of model building. Unambiguous and independent

attributes are known to incorporate precise values that make contributions to accurate prediction, which is deemed the ideal outcome. For instance, a dataset with ambiguous attributes could affect the process of model building during the training phase. Relevancy contemplation facilitates the perception of the redundant as well as irrelevant instances along with features for elimination, and it further identifies relevant entities for selection.

The chronic kidney disease (CKD) dataset, available at the University of California Irvine (UCI) repository, is considered for the present investigation, as it provides a realistic dataset that consists of both numerical and categorical variables [2]. This dataset features a number of challenging issues for researchers in model building, which include nonnormal variables, outliers, missing values, and redundant and irrelevant instances, along with class imbalance that necessitate detailed data contemplation regarding predictive analytics. This raw CKD dataset must be handled properly, as there are a few redundant and irrelevant features that may affect the performance of learning models. Further, certain instances in a dataset may be completely ambiguous, which could lead to false predictions.

Relevancy contemplation helps validate the quality of each attribute as well as its instances, and it yields readily available data for building effective machine learning models for predictive analytics. In general, feature selection algorithms deal with correlating the attributes and estimating the contribution level of each attribute toward the target variable. Features whose contribution tends toward accurate prediction of the target variable are extracted from the datasets to create an effective classification model. In predictive analytics, feature selection algorithms have been used to rank the attributes of datasets that are under consideration for various application domains [3]. The feature selection algorithms considered in this work are supervised learning algorithms used to identify irrelevant features in the CKD dataset; hence, they reduce the number of columns in a dataset. Each algorithm evaluates the attributes, assigns importance factors, and accordingly ranks and sorts the attributes. The optimum number of attributes arranged in their order of relevancy is considered inputs to various classifier models for the effective prediction of the severity of CKD.

In this work, data envelopment analysis (DEA) is conducted to identify ambiguous instances to eliminate them from the CKD dataset. Once such instances are eliminated, the optimal number of relevant features from the reduced dataset is observed with various supervised feature selection algorithms. Specifically, DEA is used for row reduction, while feature selection algorithms are used for column reduction [4, 5]. The supervised feature selection techniques considered in this work are ranked using Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) with weight optimization for further analytics; that is, the best feature selection strategy is identified using TOPSIS. The outcome of TOPSIS is validated through various classifier models to accurately predict the severity of the CKD. The classifier models are built using the reduced set of data instances given by DEA and the set of features given by each supervised feature selection algorithm. The observed results before and after detailed data contemplation and in the order of merit of various feature selection algorithms are compared and reported.

2 | STATE OF THE ART

Recently, there has been a profound degree of growth in data due to its continuous production at an ever-increasing rate in different dimensions. Likewise, data mining has become a less tangible but more challenging task. Data must be properly analyzed and tuned into the context of the given problem to make appropriate decisions and to obtain desired results. High-dimensional data deteriorate the performance of data mining methodologies and machine learning algorithms. Hence, much research has been undertaken to consider dimensionality reduction as a primary concern and exclude irrelevant, redundant, and noisy data.

In medical datasets, relevancy contemplation has been identified to be essential for simplifying the models by reducing the instances (rows) and features (columns) and avoid the curse of dimensionality and reduce training times. The raw data used for model building are analogous to crude oil, which requires further refinement before it can be used in specific applications. The outcomes of any classifier depend on the quality and relevance of the data for effective decision-making. In the context of clinical trials, statistical concepts provide guidelines for data analysis, whereas exploratory data analysis provides a wide range of strategies to derive appropriate matrices and detect anomalies or define ranges of typical values to correct input errors and impute missing values [6]. Friedrich and oters [7] reported the relevance of statistical methodology in the context of artificial intelligence (AI) development and discussed the contributions of statistics to the field of AI. This relates to methodological development, planning, and design in research studies; the assessment of data quality; data collection; and the assessment of uncertainties in the results [7]. Further, modeling concepts,

such as bias and accuracy, are also observed for evaluation in medical data analysis.

Meanwhile, Onan and Korukoğlu [8] insisted on the importance of feature selection algorithms in the development of accurate and efficient ensemble classifiers, not only in improving prediction accuracy but also in reducing learning times by considering the huge amount of data available for sentiment classification analysis. In another work, Onan [9] presented a hybrid intelligent classifier for breast cancer diagnosis with the proper identification of feature subsets and the selection of appropriate data instances. Further, Onan [10] examined the predictive performance of various ensemble classifiers in web page classification and presented a comparative analysis of four feature selection algorithms and classification models. Later, Onan [11] proposed a hybrid ensemble pruning approach to overcome the high-dimensionality problems with an identification of appropriate parametric values to improve the performance of Latent Dirichlet allocation (LDA). Here, a swarm-optimized LDA with ensemble pruning algorithm has been introduced, and its performance is tested using five biomedical text benchmarks.

Rostami and others [12] examined various swarm intelligence-based feature selection methods, mainly focused on the curse of dimensionality, and evaluated their pros and cons for general categorization. Mishra and Sharma [13] analyzed the performance of various dimensionality reduction techniques in a comparative context and showed that the LDA is more informative and accurate than the others. Musheer and others [16] analyzed various methods for the pre-processing of high-dimensional data, especially gene expression microarrays, and organized dimension reduction methods with respect to their characteristics and evaluation criteria [14]. Further, an artificial bee colony-based feature selection approach (2017, 2019) is presented to eradicate the challenges associated with independent component analysis as applied to microarray data and found the best subset of genes using the extracted features [15,16]. Independent component analysis has been used to reduce the size of the data; to optimize the reduced feature subsets, an artificial bee colony-based wrapper approach is used. The use of this hybrid approach is compared to the results obtained from the minimum redundancy maximum relevance (mRMR) method combined with the artificial bee colony algorithm for naïve Bayes (NB) classifier and with other similar biology-inspired algorithms, such as the genetic algorithm and particle swarm optimization. Zebari and oters [17] broadly analyzed a range of feature selection and extraction methodologies with a key focus on dimensionality reduction, and they identified one most accurate classifier, which featured reduced computational time.

In medical datasets, values of variables that go beyond the normal range are required for the prediction of presence or absence of a disease. However, the presence of outliers and missing values [18] can have a negative impact while building classifiers [19,20]. Points in data instances that are irrelevant for medical data analysis are handled efficiently, and instances with outliers are removed in the classifier model building process. Over the past decade, many methods of outlier detection have been reported and successfully applied across a wide range of fields, including the observation of health, credit card fraud, and detections of intrusion. Xu and others [21] have extensively studied the outlier detection methods in high-dimensional data and produced a complete understanding of outlier detection techniques. Several experiments were conducted on statistical, distance, density, clustering, deviation, and subspace-based outlier detection methods with various performance measures, including precision, average precision, AUC, rank power, and correlation coefficient, for the purpose of building an efficient classifier.

In medical data analytics, each data instance in a dataset is considered a decision-making unit (DMU). Missing values in DMUs have been effectively handled by many researchers to create accurate predictive models using machine learning algorithms. Tshering and oters [22] proposed a sequential method to identify types of missing data, including missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), for an incomplete dataset. Here, MCAR is detected using the mean and the covariance between the observed and unobserved values. Wilk's test is performed to test the null hypothesis and alternate hypotheses for the occurrence of MCAR. MAR is identified by estimating the probability of missingness in observed values and not in unobserved values, and the coefficients of logistic regression identify the missingness as MAR. MNAR uses the probability of missingness in both observed and unobserved values, and it is confirmed through a latent variable in a dataset. The authors proposed a simulation model using a Gross National Happiness dataset to validate the sequential method for each of MCAR, MAR, and MNAR data.

In this study, we focus on data instances that do not contribute much to the target attributes, as the major concern of most of the published works has been dimensionality reduction. In recent works, it was reported that the DEA is used for many application domains to identify efficient records from the datasets, including healthcare management systems [23]. However, it is hard to identify publications that report exact

identifications of efficient instances from the training datasets used for building accurate classifiers, especially in medical data analytics.

DEA is a nonparametric and linear programming technique used to estimate the efficiency of each data instance by observing its relative performance [24]. DEA identifies instances that require imputation as infinite efficient records and thus facilitates the handling of missing values in those instances. Although DEA does not identify outliers, it provides an efficient means of identifying inefficient instances in medical datasets. Thus, it enables row reduction, while feature selection algorithms facilitate column reduction by identifying irrelevant features [25,26]. Moreover, it is well known that certain data instances contribute less to a target, which leads to a lower classifier accuracy. The performance measures of the classifiers as well as the feature selection algorithms are assessed for their accuracy in their prediction, not only after imputation and removal of outliers but also following the removal of inefficient instances. The models developed so far in medical data analytics focus only on the enhancement of feature selection algorithms and integrated classifiers to improve prediction accuracy. Hence, the major concern of most research work is column reduction rather than row reduction.

Wibawa and others [27] developed a machine learning model using ensemble learning and feature selection to enhance the quality of CKD diagnosis. The correlation-based feature selection (CFS) algorithm and AdaBoost have been used herein for ensemble learning to improve the detection of the presence of CKD. The K-nearest neighbor (KNN) algorithm, NB, and support vector machine (SVM) are assumed as base classifiers. Using them, it was proven that the best result was achieved through the combination of KNN classifier with CFS and AdaBoost, producing a 0.981 accuracy rate, a 0.980 recall rate, and a 0.980 f-measure rate. Grissa and others [28] proposed a workflow that describes the general feature selection process, using appropriate methodologies for predictive biomarker discovery. Their study focused on the machine learning methods, namely, SVM-recursive feature elimination (RFE), random forest (RF), and RF-RFE, as well as on univariate statistical analysis of variance (ANOVA), and a comparative study was performed on an original metabolomics dataset with reduced subsets. The relevant features were extracted from the combination of these different methods using importance scores. From the results obtained, the RF-Gini method combined with ANOVA was identified as the best one for feature selection for the early prediction of biomarkers. A classifier model was built using linear logistic regression on this reduced dataset to identify the top five attributes, and it was validated in terms of prediction accuracy.

Qin and others [29] proposed an integrated classifier model that combines LOG and RF by perceptron over the CKD dataset to improve prediction accuracy. Initially, the dataset was tuned using KNN imputation, where the numerical missing values are filled by the median, and the categorical missing values are filled with the mode of K-samples. Further, LOG, RF, SVM, KNN, NB, and feedforward neural network classifiers are evaluated using the complete and tuned CKD dataset for optimal feature selection. The experimental results show that the integrated RF model with Signum activation function has good performance for the CKD diagnosis.

Chen and others [1] utilized three datasets with a higher number of variables, namely, a bank marketing database, a car evaluation database, and human activity recognition using smartphones to perform an analysis of appropriate feature selection. Here, the accuracy and performance of classifiers such as RF, SVM, KNN, and LDA are evaluated, wherein RF emerges as an efficient and accurate algorithm in terms of accuracy. Moreover, all of the classifiers are trained across raw datasets, and such datasets are subjected to the feature selection methods RFE and Boruta.

The importance of feature selection algorithms in model building has been studied for decades, but only a few works have ordered or ranked the algorithms by considering their efficiency. The TOPSIS with weight optimization has been proposed in this work to rank the feature selection algorithms. TOPSIS is a multi-criterion decision analysis method for ranking different alternatives based on various criteria. More recently, multiple criteria decision-making (MCDM) methods have been used to rank and evaluate the performance of the features. TOPSIS is unique relative to MCDM methods because it depends on logical thinking, which is based on the simultaneous evaluation of the nearest distance from the best alternative (positive ideal solution) and the longest distance from the worst alternative (negative ideal solution). TOPSIS also incorporates a straightforward approach, which is suitable for cases with a larger or smaller numbers of criteria, and it is appropriate to use with qualitative or quantitative data. TOPSIS has been used to obtain an overall performance value for each alternative to make a final decision. Esfandiari and Rizvandi [30] have adopted the TOPSIS technique to rank business development strategies, including critical success factors analysis, business systems planning, Porter's forces model, SWOT analysis, value chain analysis, and MIN. The major challenge of TOPSIS implementation in medical data analytics is the selection of weighted vectors, which should be framed in consultation with medical practitioners in respective fields.

In this work, DEA is carried out on the CKD dataset to identify efficient instances. Before the removal of inefficient instances by DEA according to their relative efficiency, the dataset was appropriately imputed, followed by the removal of instances consisting of outliers. The remainder of this article includes the following sections to address all of the objectives, as follows: (i) Section 3 reveals the proposed methodology, including data cleaning, identification of inefficient instances, selection of relevant features, and ranking of feature selection algorithms, (ii) Section 4 validates the ranked feature selection algorithms using prominent classifiers, and (iii) Section 5 summarizes the results and discussion.

3 | PROPOSED METHODOLOGY

In the medical data analytics, the major concern is dimensionality reduction, but the work addressed herein is focused on data instances, which do not contribute much to the target attribute. To build efficient classifiers, the data instances that are more inclined toward the target attribute must be identified for accurate prediction of different stages of CKD. The most relevant features should be selected from the set of feature selection algorithms, which are ranked based on their order of efficiency. Relevant data instances are identified by DEA, while relevant features are estimated by their contribution toward the target variable. The feature selection algorithms, which are used to grade the features and identify the best among them, are ranked using weight-optimized TOPSIS. The ranked feature selection algorithms are validated through classifier models used to produce accurate prediction of the severity of the disease. Rows are reduced by removing inefficient data instances through DEA, whereas the columns are reduced by selecting relevant features in the feature selection algorithms.

3.1 | Data collection

The CKD dataset for this present work is obtained from UCI Machine Learning Repository. It has 400 instances and 24 independent variables, with 11 quantitative and 13 qualitative variables to predict the disease as ckd or notckd, which is a binomial response. The dataset consists of clinical test records of real cases that are considered as data instances in the collection for CKD predictive analytics. The quantitative independent variables are continuous and discrete, while the qualitative variables are nominal and ordinal.

3.2 | Feature extraction for multinomial response

Extracting relevant features enhances the analytics process to identify new patterns or to change the nature of the responsiveness of the existing target variable. The new feature, that is, estimated glomerular filtration rate (eGFR), is extracted using the patient's demographic information and clinical reports to predict the presence of CKD and its stages as per the guidelines of Kidney Disease Improving Global Outcomes (KDIGO) [31]. eGFR is estimated using 1, which is known as the modification of diet in renal disease equation. eGFR is extracted with the help of the attributes "age" and "sc" and populated demographic attributes "race" and "gender."

$$\mathrm{eGFR} = 175 \times (\mathrm{sc})^{-1.154} \times (\mathrm{age})^{-0.203} \times \mathrm{gender\_condition} \times \mathrm{race\_condition}, \quad (1)$$

where $\mathrm{gender\_condition} = \begin{cases} 0.742, \text{if female} \\ 1, \text{if male} \end{cases}$ and $\mathrm{race\_condition} = \begin{cases} 1.212, \text{if black} \\ 1, \text{if others} \end{cases}$. The "class" of each instance of the enhanced CKD dataset is labeled using one of the values from the multinomial responsive set {stages 1–5} against the binomial response {notckd, ckd}, as per KDIGO guidelines. The enhanced CKD dataset consists of 28 attributes, with the "class" attribute changed from binomial to multinomial target; hence, there are 27 independent variables. In the enhanced CKD dataset, the distribution of data instances among various stages is represented as follows: stage 1, 25.25%; stage 2, 19.5%; stage 3, 22.75%; stage 4, 14.75%; and stage 5, 17.75%.

3.3 | Data pre-processing

Data pre-processing is the preliminary process in data analysis. Data cleaning is carried out by handling outliers and missing data, taking account of the distribution, types of variable, and types of missingness. The enhanced CKD dataset has missing values in 242 instances, and the numerical independent variables are highly skewed. The skewness may be caused by missing values or the presence of outliers; hence, such issues are addressed. An enhanced dataset with additional extracted features is used for data tuning following a stratified split. Splitting the enhanced CKD dataset randomly into training and test datasets leads to an imbalanced distribution of classes (i.e., CKD stages), which could affect the performance of models for the minority classes [32,33]. Therefore, the

dataset was split in a stratified way (in the ratio of 70:30) to maintain an equal distribution of classes. The training dataset (70%) and the test dataset (30%) were pre-processed separately to prevent data leakage. If this is not done, the prediction accuracy in the production environment is reduced, as the information outside of the training dataset might have been used for building the model. Therefore, the fine-tuned training dataset is used by the hybrid ensemble, and similarity-based and probabilistic classifiers are adopted to learn, and the pre-processed test dataset is utilized to validate the learned models.

3.4 | Outlier detection

The main objective of outlier detection is to extract the outliers of all numeric variables, as the ordinal variables in the CKD dataset do not contribute to the outliers. A hybrid approach is introduced to handle the outliers present in the dataset. The interquartile range (IQR) and mean are integrated using the skewness of the attributes to handle the outliers. The third standard deviation (STD) from mean of each numerical attribute (Z-score) is estimated and integrated with the boxplot using the IQR to detect outliers from the CKD dataset. The data instances with variables beyond their upper and lower threshold ranges of values are identified as outliers and removed.

All far points are not treated as outliers, as their presence is highly useful for accurately predicting the disease, so it is recommended to exclude only the extreme far points. In the statistical analysis, the boxplot (IQR) detects 138 records, that is, 34.5% in the CKD dataset as outliers, and the Z-score (the third STD from the mean) detects 41 records, that is, 10.25% as outliers, and the proposed hybrid method detects 1% of records, that is, 4 records, as outliers.

3.5 | Handling missing values

Missing values are commonly attributed to human error. Missingness takes three forms. MAR is visualized through the scatterplot by showing the relationship between two variables and is imputed by the mean-median-mode. MNAR is imputed by the machine learning model KNN. Then, MCAR is detected using a list-wise method and imputed by the mean-median-mode. If the variables' correlation coefficient is 1, they are grouped as MNAR, as they exhibit the same missing pattern. If their correlation coefficient is 0, they are grouped as MCAR, as they do not exhibit the same missing pattern. Lastly, if it is less than 1, they are grouped as MAR, as a relationship exists among the variables.

The sets of variables identified to have missing values due to of MNAR are {*pcc*, *ba*}, {*sod*, *pot*}, {*htn*, *dm*, *cad*}, {*appet*, *pe*, *ane*}, and {*gfr*, *race*, *gender*}, and the set of variables with missing values due to MCAR is {*rbc*, *pc*, *bgr*, *age*, *bp*}, while the sets due to MAR are {*wc*, *rc*}, {*hemo*, *pcv*}, {*bu*, *sc*}, and {*al*, *su*, *sg*}. The MNAR type of missing values are imputed by *K*NN with $K = 5$. The MCAR and MAR types of missing values are imputed using the mean-median-mode. After data cleaning, the inefficient data instances are identified, and the features that influence the prediction are recognized as building the classifiers for validation.

3.6 | Data envelopment analysis

DEA identifies ambiguous instances for elimination. Each data instance is a DMU, and its efficiency is measured using DEA. The role of DEA in data analytics is not only intended for dimensionality reduction but also for the validation of data cleaning; that is, it confirms the instances handled to set right the missing values during data pre-processing. The importance of relevancy examination in data analytics is proliferated by investigating the contributions of both independent variables and of data instances, with respect to the target attribute and the model's efficiency.

A stratified split of the entire dataset yields 281 data instances in the training dataset and 119 in the test dataset. After the split, the numbers of instances for the stages in the training dataset are about 71, 55, 64, 40, and 51. The numbers of data instances in the training dataset are 277 after the detection of outliers (as discussed in Section 3.4), leaving 71, 55, 64, 40, and 47 instances in the different stages; that is, there are no cases of outliers in the instances for stages 1–4.

Imputation is carried over to handle the missing values in these 277 data instances (Section 3.5). Further, nine data instances are identified as being inefficient, as stated by Algorithm 1. Hence, 268 data instances from the training dataset are marked as efficient instances, and they are used for feature selection and model building. After inconsistent data instances were removed from the training dataset by DEA, the number of instances for stages 1–5 is computed as 69, 52, 61, 39, and 47, respectively. From these results, it can be observed that DEA does not remove any instance of stage 5 once the outliers are handled properly. As DEA only handles numerical data, the categorical variables in the training dataset

are converted into numeric quantities through label encoding.

DEA Algorithm

Inputs: Data instances—277 and attributes—{*cad, htn, dm, pcc, ba, ane, appet, pe, sod, pot, age, bp, sg, al, su, rbc, pc, bgr, bu, sc, hemo, pcv, wc, rc, gender, race, gfr*}
Output: {*class*}

Step 1: Construct the normalized decision matrix using the following equation.

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{n=1}^{i} x_{ij}^2}},$$

i is the number of data instances considered for analysis, $j = 1,..., 27$

Step 2: The efficiency (h_i) for individual DMU is defined as:

$$h_i = \text{Max}\left(\frac{\sum_{q=1} v_q y_{qi} - y_i}{\sum_{p=1}^{27} u_p x_{pi}}\right)$$

where,

- y_{qi} is the value of qth output for ith unit
- x_{pm} is the value of pth input for ith unit
- u_p and v_q are nonnegative values
- v_q is the weight associated with pth output
- u_p is the weight associated with pth input
- i DMUs have been compared based on 27 inputs and 1 output

when same weights $\{v_q,\ u_p\}$ are applied to all DMUs, the efficiency h_i should not be greater than 1, that is,

$$\text{Max}\left(\frac{\sum_{q=1} v_q y_{qi} - y_i}{\sum_{p=1}^{27} u_p x_{pi}}\right) \text{ should be} \leq 1$$

The above constraint is represented as:

$$\sum_{q=1} v_q y_{qi} - y_i - \sum_{p=1}^{27} u_p x_{pi} \leq 0$$

DEA method solves n linear programming problems, and it cannot handle fraction; hence, the formulation shown in Step 2 of the above algorithm has been transformed as follows:

$$g_n = \text{Max}\left(\sum_{q=1} v_q y_{qi} - y_i\right)$$

such that

$$\sum_{p=1}^{27} u_p x_{pi} = 1$$

where $v_q \geq 0$ and $u_p \geq 0$.

Output: Efficient data instances: 268

3.6.1 | Relevancy analysis of CKD dataset using DEA—Insightful observations

The DMUs with the lowest input and highest outputs are considered as efficient units. The DEA algorithm is implemented using the R programming language to identify data instances with missing values with infinite efficiency. Of the 400 data points, 242 data instances have missing values, identified as infinitely efficient DMUs, with a relative efficiency of less than 100% for 106 data instances. The identified inefficient DMUs are made efficient through the proportional reduction of their inputs, while the production of their outputs is held constant. The insights obtained through DEA in the CKD dataset to identify the efficient data instances are given in Table 1. The CKD dataset is subjected to DEA before imputation, but only 13%, that is, 52 data instances, were identified as efficient instances.

The dataset is split in a stratified manner, and 281 data instances of the training dataset are subjected to DEA before and after imputation. DEA identifies 214 and 9 data instances as inefficient before and after imputation, respectively. After handling the outliers from the training dataset, 4 instances are removed, and 277 data instances are subjected to DEA. Herein, 209 instances are identified as inefficient before imputation, and 9 instances are observed to be inefficient after imputation. Hence, DEA yields 268 instances from the training dataset for further processing of feature selection and model building. After imputation and before outlier removal from the training dataset, the data instances with outliers are not considered inefficient instances by DEA. This is so because the data instances with outliers possess distant data points at the maximum end, which then increases relative efficiency.

3.7 | Ranking of feature selection algorithms for medical data analytics

Removing redundant and irrelevant features during model building with machine learning algorithms has equal importance to the selection of relevant features. It is focused on deriving the best strategy to select an appropriate feature selection algorithm, which can be suitably adopted for any medical dataset in place of finding the best features with accuracy. While analyzing the CKD dataset, ranking among the most frequently used feature selection methods is performed using TOPSIS, such as extra-trees classifier, analysis of variance, RFE, chi-squared test, and MI, which is enhanced with weight optimization.

3.8 | TOPSIS-based ranking of feature selection algorithms

The stated feature selection methods are ranked using TOPSIS for 281, 277, and 268 data instances of the training dataset, that is, (i) the entire set of data instances (training dataset, 281 instances), (ii) the set of data

TABLE 1 Descriptive statistics of DEA on CKD dataset

Sr. No.	Data instances category	No. of data instances	No. of instances with infinite efficiency	No. of inefficient instances	No. of efficient instances
1	Raw CKD dataset	400	242	348	52
2	70% raw training dataset	281	172	214	67
3	70% training dataset with imputation	281	0	9	272
4	70% training dataset with outliers handled	277	169	209	68
5	70% training dataset with imputation and outliers handled	277	0	9	268

TABLE 2 Evaluation matrix with ranking of each attribute (268 data instances)

Algorithms	cad 1	htn 2	hemo 3	bu 4	sc 5	...	al 26	egfr 27
ETC	25	9	15	26	27	...	16	1
ANOVA	15	10	13	16	20	...	25	1
RFE	24	10	16	22	26	...	25	2
Chi^2	15	10	13	16	20	...	25	1
MI	25	7	13	26	23	...	19	1

instances in the training dataset after removal of outliers (277), and (iii) the set of data instances in the training dataset after removal of inefficient data instances by DEA (268). In each case, the ranks of all of the attributes derived from each feature selection algorithm are inputted to the proposed TOPSIS model by way of constructing the evaluation matrices. The evaluation matrix is constructed using the inputs from 268 data instances, as shown in Table 2. While each feature selection algorithm ranks the attributes, TOPSIS ranks the algorithm itself. The aim is to measure its performance (by estimating the t score in TOPSIS) of each feature selection algorithm, before and after the handling of outliers and the removal of inefficient instances by DEA. It has been proven that the t-score values of the top-ranked feature selection algorithms are high, only after the data have been completely pruned.

The feature selection algorithms are evaluated using all of the attributes in which the eGFR is the most important criterion and the attributes bacteria (ba), gender, race, and coronary artery disease (cad) are observed to have the least importance. The ranking of feature selection algorithms is evaluated using the TOPSIS score (t score), which is based on the similarity to the worst alternative. The step-by-step procedure for the TOPSIS-based ranking, with enhanced weight optimization using the CKD training dataset D with v (=27) variables and N data instances (=281, 277, and 268 records) and m (=5) feature selection algorithms with ranked variables as input is depicted in the form of a flow chart in Figure 1.

The ranks of the attributes given in the evaluation matrices are normalized for all cases. The weight vector in TOPSIS is generally assigned by experts based on the problem domain, and in the case of CKD dataset, the expert's opinion is that no clinical factor can be given as a weight for the cause of the disease. Therefore, the degree to which each attribute fits into the normal distribution as estimated by Shapiro–Wilk test is taken in place of a weight, and these factors are normalized. Weighted normalized decision matrices are generated, and the minimum and maximum value of each variable in its column is estimated. Further, the Euclidean distance between the target alternative and the best/worst alternative for all cases is calculated. Each algorithm ranks features according to the estimated scores and treats the category of each feature as a cost. Hence, the minimum and maximum of each attribute are considered as the best and worst alternatives, respectively.

The best feature selection algorithm is identified by computing the similarity and its closeness to the worst

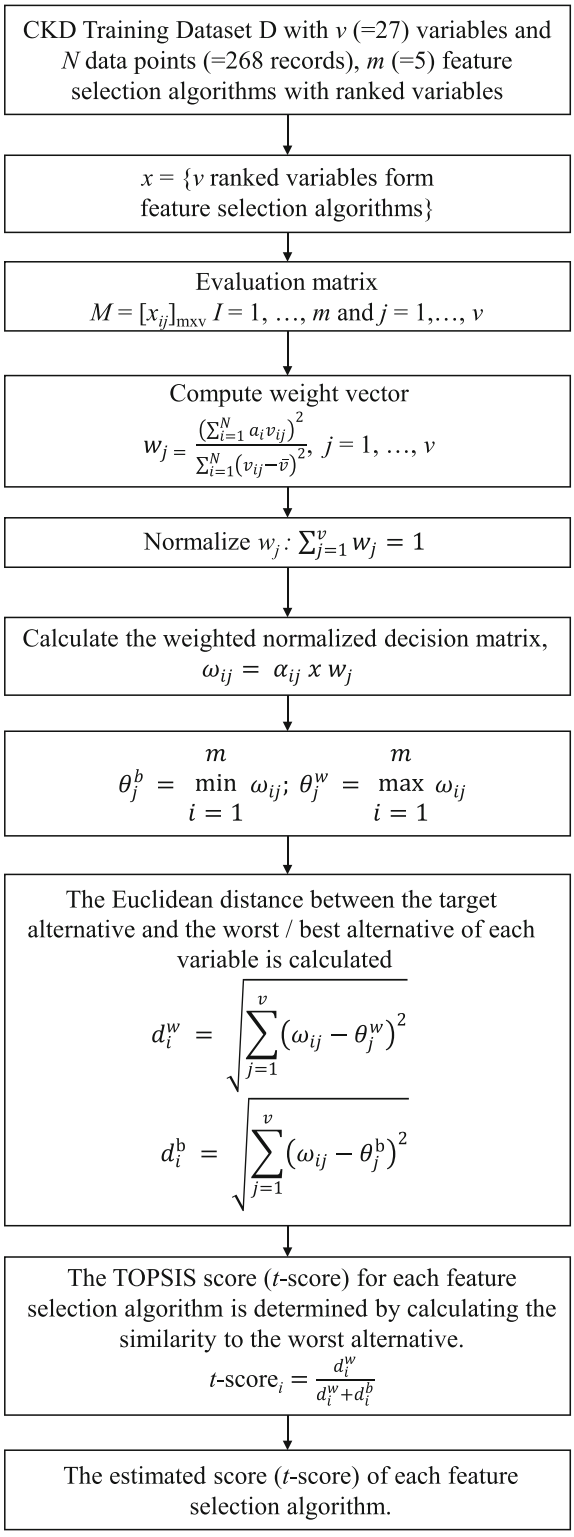

FIGURE 1 Ranking of feature selection algorithms using TOPSIS

alternative (t score). The t-score values of each feature selection algorithm by considering 281, 277, and 268 data instances are listed in Table 3.

TABLE 3 T score for each feature selection algorithm

Feature selection algorithm	Data instances		
	268	277	281
ETC	0.6564	0.6371	0.7045
MI	0.6191	0.5733	0.6844
ANOVA	0.5533	0.5638	0.5919
RFE	0.5345	0.5368	0.4118
Chi^2	0.5212	0.5139	0.5794

TABLE 4 Ranking of feature selection algorithms

Data instances	Ordered feature selection algorithms
281	ETC > MI > ANOVA > Chi^2 > RFE
277	ETC > MI > ANOVA > RFE > Chi^2
268	ETC > MI > ANOVA > RFE > Chi^2

The feature selection algorithm extra-trees classifier is ranked highly by the weight-optimized TOPSIS method, with the t scores of 0.6564, 0.6371, and 0.7045 while considering 268, 277, and 281 data instances, respectively. Based on the estimated t scores, the feature selection algorithms are ranked for all cases and are presented in Table 4 in their order of merit.

4 | VALIDATION

The feature selection algorithms such as extra tree classifier (ETC), MI, ANOVA, RFE, and chi-squared test rank the attributes of the CKD dataset based on their relevancy. The top 10 attributes in their order of merit are identified by the feature selection algorithms, which are more relevant for the target variable to facilitate the accurate prediction analysis, as listed in Table 5.

ETC is ranked at the top, and the best features selected are gfr, sc, bu, hemo, gender, age, sod, bgr, htn, and bp. The best features identified by each feature selection algorithm are validated for their truth using the hybrid ensemble and similarity-based classifiers and further compared with classifiers such as random forest, AdaBoost, NB, KNN, and SVM.

The 10 top-ranked relevant features, extracted in the order of their merit by the feature selection algorithm, are given as inputs to build up the models. To estimate the prediction accuracy relative to the actual extracted target value in the test dataset, built-up models are used. To validate the ranked feature selection algorithms precisely for all cases and accurately predict the severity of the disease, two new classifier models are developed. The first model groups multiple classifiers to improve

prediction accuracy with proper assignment of weights via a voting process (hybrid ensemble classifier). In this model, a hybrid strategy is adopted by integrating bagging (random forest) and boosting (AdaBoost) techniques through hard voting classifiers for the appropriate prediction of the CKD stage with tenfold cross-validation to classify new data instances correctly by reducing bias and variance [34, 35].

TABLE 5 Relevant attributes for prediction analysis (268 data instances)

Feature selection algorithm	Top 10 attributes in the order of relevance
ETC	*gfr*, *sc*, *bu*, *hemo*, *gender*, *age*, *sod*, *bgr*, *htn*, *bp*
MI	*gfr*, *sc*, *bu*, *hemo*, *pcv*, *rc*, *htn*, *sod*, *pot*, *bgr*
ANOVA	*gfr*, *sc*, *bu*, *hemo*, *htn*, *pcv*, *rc*, *sod*, *dm*, *age*
RFE	*sc*, *gfr*, *age*, *bp*, *pcv*, *sod*, *al*, *hemo*, *rc*, *htn*
Chi^2	*gfr*, *bu*, *wc*, *bgr*, *sc*, *age*, *pcv*, *al*, *bp*, *htn*

The second built-up model (weighted decision matrix classifier) is a knowledge-based system that uses historically similar cases to interpret new unseen data points. In this model, the data instances are segmented into independent decision matrices, and each segment has its own mean and distribution along with a weighted DMU. The reduced feature set used by this classifier is based only on the predictors' coefficients, and the variability exhibited by them is not captured. Each weighted DMU estimates the decision by calculating the sum of product of normalized weight of each decision variable and the difference between the test data and the predictors' mean of decision matrix. The *k*th weighted DMU, which yields the minimum value, is the actual predictor of the severity of the disease.

The feature selection algorithms which are ranked by weight-optimized TOPSIS are validated through the above built-up models for accurate prediction of severity of CKD. The performance matrices of accuracy, precision, recall, and F1-score are compared with probabilistic NB classifier, decision tree-based random forest, and AdaBoost with KNN and SVM classifiers. The accuracy validation of the ranked feature selection

TABLE 6 Validation of feature selection algorithms

	Accuracy (%)						
Feature selection algorithm	**Integrated ensemble classifier**	**WDM classifier**	**Random forest**	**Ada boost**	**Naïve Bayes**	**KNN**	**SVM**
(a) **281 data instances**							
ETC	98.31	94.61	97.4	96.6	82.35	75.63	88.23
MI	97.47	93.94	94.9	96.6	85.71	84.87	89.91
ANOVA	97.47	92.93	94.1	96.6	81.51	86.5	92.43
RFE	97.47	95.28	94.95	93.2	86.55	95.79	95.79
Chi^2	97.47	92.26	95.79	96.6	84.87	33.61	87.39
(b) **277 data instances**							
ETC	98.31	92.26	96.63	94.11	83.19	78.15	83.19
MI	98.31	86.71	94.11	89.91	83.19	77.31	90.75
ANOVA	98.31	89.16	94.95	86.55	84.03	32.77	83.19
RFE	96.63	93.44	94.95	90.75	85.71	86.55	81.51
Chi^2	98.31	91.54	95.79	82.35	84.03	32.77	83.19
(c) **268 data instances**							
ETC	98.31	93.94	94.95	97.47	82.35	78.15	82.35
MI	98.31	93.93	95.79	97.47	83.19	78.9	89.07
ANOVA	98.31	93.60	94.95	95.79	84.03	84.03	89.91
RFE	97.47	94.61	94.95	96.63	85.71	86.55	85.71
Chi^2	97.47	91.93	95.79	96.63	84.87	33.61	87.39

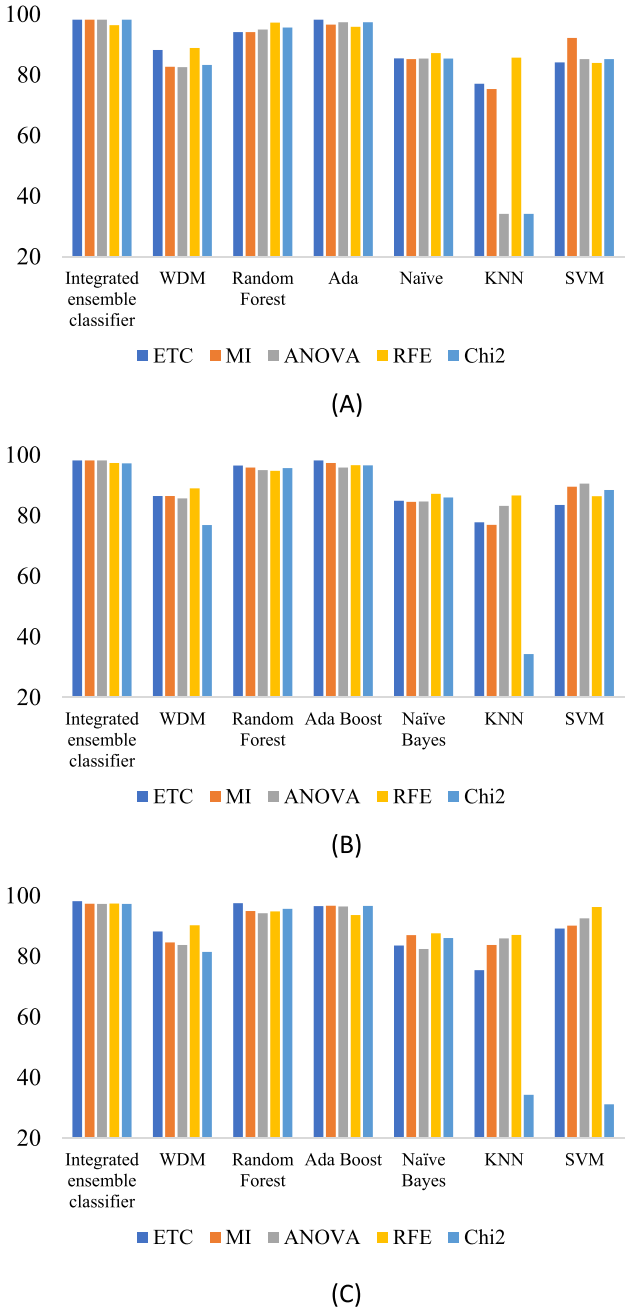

FIGURE 2 (A) Precision—268 data instances; (B) precision—277 data instances; (C) precision—281 data instances

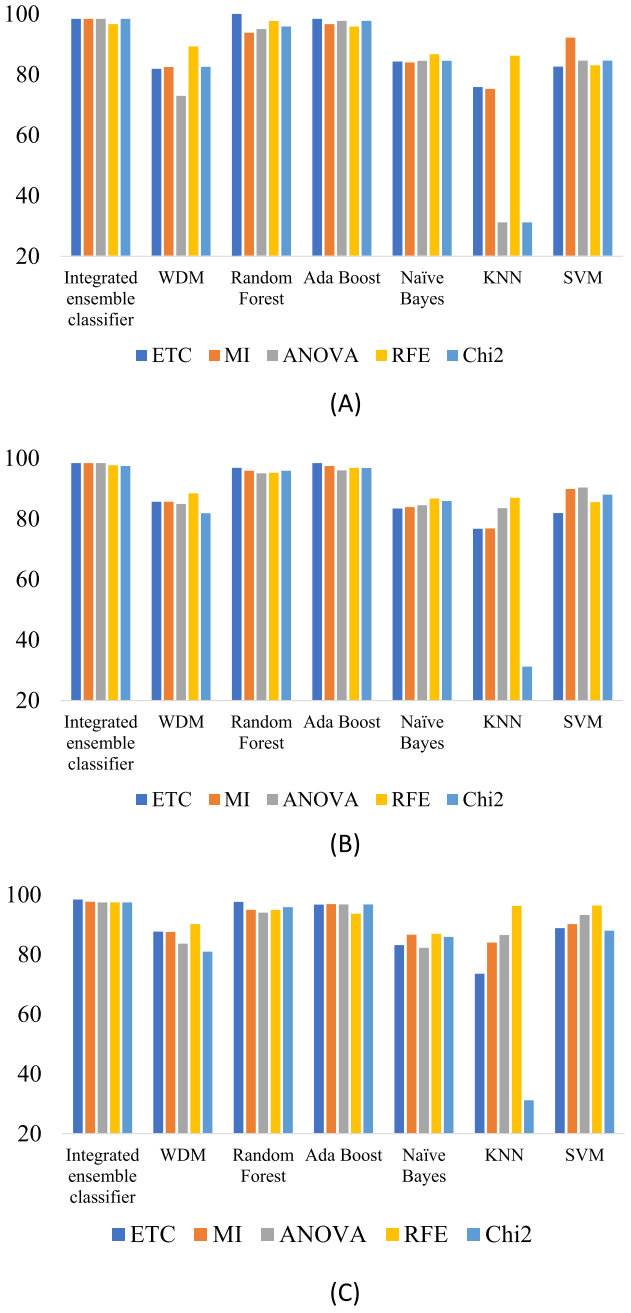

FIGURE 3 (A) Recall—268 data instances; (B) recall—271 data instances; (C) recall—281 data instances

algorithms for all cases, taking into account 281, 277, and 268 data instances, is given in Table 6 (a–c) respectively. The other quality parameters of precision, recall, and F1-score are displayed in Figures 2–4, respectively.

It can be observed that the top-ranked 10 relevant attributes extracted from the extra-trees classifier yield better test accuracy, that is, a consistent rate of 98.31% in all cases, while given as input to the integrated ensemble classifier. When the integrated ensemble model is built after the removal of outliers (with 277 instances) and inefficient instances according to DEA (268 instances) by adopting the feature selection algorithms MI and ANOVA, it achieves the same test accuracy, that is, 98.31%. The precision, recall, and F1-score measures are high for the features given by ETC, MI, and ANOVA. Comparing the results presented in Table 6 (a–c), it is strongly recommended to utilize

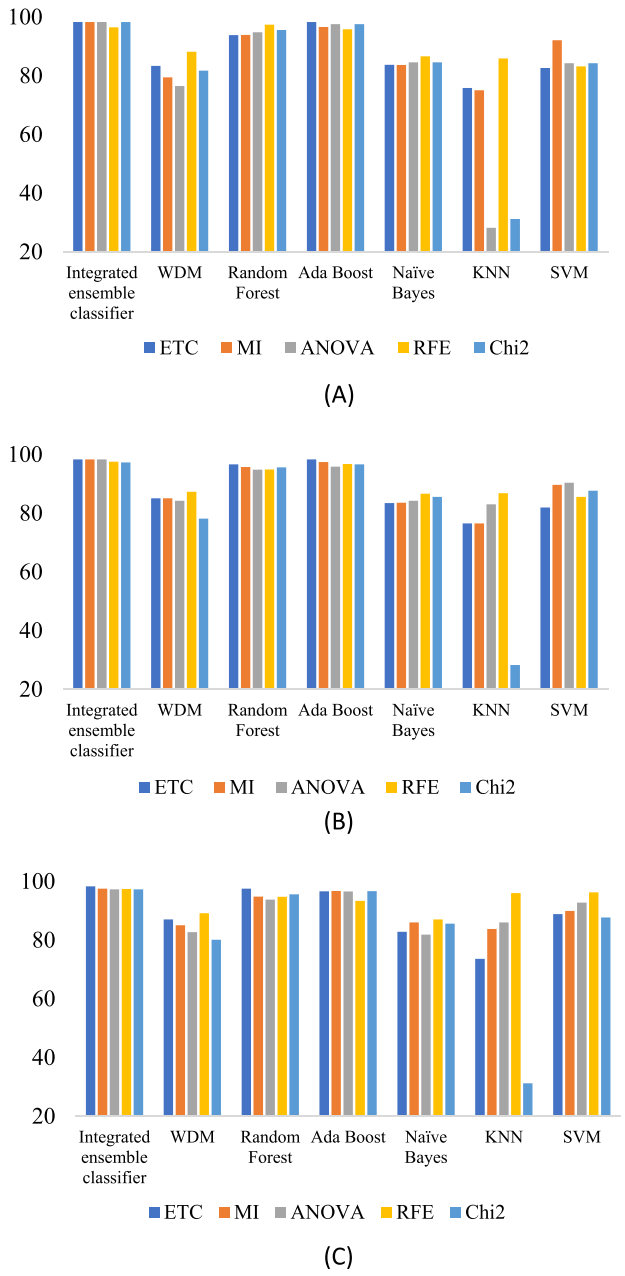

FIGURE 4 (A) F1-score—268 data instances; (B) F1-score—277 data instances; (C) F1-score—281 data instances

the feature selection algorithms ETC, MI, and ANOVA in the same order. Further, the hybrid ensemble, WDM, RF, and AdaBoost classifiers, are preferred to the NB, KNN, and SVM classifiers to accurately predict the severity of the disease.

5 | CONCLUSIONS

The raw CKD dataset has been considered for analysis, and the statistical insights of the attributes with their importance for envisaging the CKD are presented. The feature eGFR is extracted to predict the severity of the disease. The outliers are detected using deep insightful observations through statistical analysis after the implementation of a new hybrid model with Z score and IQR. The issue of data leakage is addressed by pre-processing the training and test datasets separately. The enhanced dataset is then split into training and test datasets in a stratified manner in a 70:30 ratio to balance the class distribution. The impact of DEA on the complete training dataset before and after imputation and the removal of outliers are analyzed, and the CKD dataset is then perfectly tuned. A set of five supervised feature selection algorithms are ranked by TOPSIS, which is enhanced with weight optimization; then, the relevant features are selected by taking into account all of the cases for model building. The best feature selection algorithm is identified by computing the similarity and its closeness to the worst alternative and the TOPSIS scores (t scores) of the feature selection algorithms in the order of merit, as follows:

$$\{(\text{ETC}, 0.6564), (\text{MI}, 0.6191), (\text{ANOVA}, \ 0.5533), (\text{RFE}, 0.5345), (\text{Chi2}, \ 0.5212)\}.$$

A hybrid classifier is designed for the prediction of the severity of CKD by integrating boosting and bagging techniques through the voting classifier. The WDM classifier is built by considering the instance similarities and relationships among the variables. The ranked feature selection algorithms are validated through effective classifier models, which include hybrid ensemble classifier and WDM classifier toward accurate prediction of the severity of the CKD. The feature selection algorithms ETC, MI, and ANOVA have been determined to outperform in all cases in terms of accuracy; using hybrid ensemble classifier and with respect to WDM classifier, the RFE algorithm is noted to perform well.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest.

ORCID

P. Antony Seba https://orcid.org/0000-0002-3209-6858

REFERENCES

1. R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, *Selecting critical features for data classification based on machine learning methods*, J Big Data **7** (2020), no. 52.

2. UCI, *Chronic kidney disease data set*. Available at: https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease [last accessed February 2020]
3. Z. M. Hira and D. F. Gillies, *A review of feature selection and feature extraction methods applied on microarray data*, Adv. Bioinform. **2015** (2015), 198363.
4. S. Rousseau and R. Rousseau, *Data envelopment analysis as a tool for constructing scientometric indicators*, Scientometrics. **40** (1997), 45–56.
5. N. Cui, J. Hu, and F. Liang, *Complementary dimension reduction*, Stat. Anal. Data. Min. **14** (2020), 1–10.
6. B. M. Konopka, F. Lwow, M. Owczarz, and L. Laczmanski, *Exploratory data analysis of a clinical study group: Development of a procedure for exploring multidimensional data*, PLOS One **13** (2018), e0201950.
7. S. Friedrich, G. Antes, S. Behr, H. Binder, W. Brannath, F. Dumpert, K. Ickstadt, H. A. Kestler, J. Lederer, H. Leitgöb, M. Pauly, A. Steland, A. Wilhelm, and T. Friede, *Is there a role for statistics in artificial intelligence?* Adv. Data Anal. Classif. (2021). https://doi.org/10.1007/s11634-021-00455-6
8. A. Onan and S. Korukoğlu, *A feature selection model based on genetic rank aggregation for text sentiment classification*, J. Inf. Sci. **43** (2017), 25–38.
9. A. Onan, *A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer*, Expert Syst. Appl. **42** (2015), 6844–6852.
10. A. Onan, *Classifier and feature set ensembles for web page classification*, J. Inf. Sci. **42** (2016), 150–165.
11. A. Onan, *Biomedical text categorization based on ensemble pruning and optimized topic modelling*, Comp. Math. Methods Med. **2018** (2018), 2497471.
12. M. Rostami, K. Berahmand, E. Nasiri, and S. Forouzande, *Review of swarm intelligence-based feature selection methods*, Eng. Appl. Artif. Intel. **100** (2021). https://doi.org/10.1016/j.engappai.2021.104210
13. D. Mishra and S. Sharma, *Performance analysis of dimensionality reduction techniques: A comprehensive review*, in *Advances in Mechanical Engineering. Lecture Notes in Mechanical Engineering*, Springer, Singapore, 2021, pp. 639–651.
14. R. Aziz, C. K. Verma, and N. Srivastava, *Dimension reduction methods for microarray data: A review*, AIMS Bioeng. **4** (2017), 179–197.
15. R. Aziz, C. K. Verma, and N. Srivastava, *A novel approach for dimension reduction of microarray*, Comput Biol. Chem. **71** (2017), 161–169.
16. R. A. Musheer, C. Verma, and N. Srivastava, *Novel machine learning approach for classification of high-dimensional microarray data*, Soft Comput. **23** (2019), 13409–13421.
17. R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, *A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction*, J. Appl. Sci. Technol. Trends. **1** (2020), 56–70.
18. W. Young, G. Weckman, and W. Holland, *A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits*, Theor. Issues Ergon. Sci. **12** (2011), 15–43.
19. J. Tang, Z. Chen, A. W. Fu, and D. W. Cheung, *Capabilities of outlier detection schemes in large datasets, framework and methodologies*, Knowl. Inf. Syst. **11** (2006), 45–84.
20. S. Thudumu, P. Branch, J. Jin, and J. Singh, *A comprehensive survey of anomaly detection techniques for high dimensional big data*, J. Big Data **7** (2020).
21. X. Xu, H. Liu, L. Li, and M. Yao, *A comparison of outlier detection techniques for high-dimensional data*, Int. J. Comp. Intell. Syst. **11** (2018), 652–662.
22. S. Tshering, T. Okazaki, and S. Endo, *A method to identify missing data mechanism in incomplete dataset*, Int. J. Comput. Sci. Network Sec. **13** (2013).
23. Y. A. Ozcan and K. Tone, *Health care benchmarking and performance evaluation: An assessment using data envelopment analysis (DEA)*, 2nd ed., Springer, New York, NY, USA, 2014.
24. E. Thanassoulis, K. D. Witte, J. Johnes, G. Johnes, G. Karagianni, and C. S. Portela, *Applications of data envelopment analysis in education*, In *Data envelopment analysis. International series in operations research & management science*, Vol. **238**, Springer, Boston, MA, USA, 2016.
25. D. Jain and V. Singh, *Feature selection and classification systems for chronic disease prediction: A review*, Egypt Inform. J. **19** (2018), 179–189.
26. X. Wang, Y. Yan, and X. Ma, *Feature selection method based on differential correlation information entropy*, Neural Process Lett. **52** (2020), 1339–1358.
27. M. S. Wibawa, I. M. D. Maysanjaya, and I. M. A. W. Putra, Boosted classifier and features selection for enhancing chronic kidney disease diagnose, (International Conference on Cyber and IT Service Management, Denpasar, Indonesia), Aug. 2017. https://doi.org/10.1109/CITSM.2017.8089245
28. D. Grissa, M. Pétéra, M. Brandolini, A. Napoli, B. Comte, and E. P. Guillot, *Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data*, Front. Mol. Biosci. **3** (2016). https://doi.org/10.3389/fmolb.2016.00030
29. J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, *A machine learning methodology for diagnosing chronic kidney disease*, IEEE Access. **8** (2020), 20991–21002.
30. M. Esfandiari and M. Rizvandi, *An application of TOPSIS method for ranking different strategic planning methodology*, Manag. Sci. Lett. **4** (2014), 1445–1448.
31. K DIGO, *KDIGO clinical practice guideline on the ecaluation and management of candidates for kidney transplantation*, 2018.Available at: https://kdigo.org/wp-content/uploads/2018/08/KDIGO-Txp-Candidate-GL-Public-Review-Draft-Oct-22.pdf
32. N. Saravanan, G. Sathish, and J. M. Balajee, *Data wrangling and data leakage in machine learning for healthcare*, J. Emerg. Technol. Innov. Res. **5** (2018), https://www.jetir.org/papers/JETIRC006413.pdf
33. F. Farias, T. Ludermir, and C. B. Filho, *Similarity based stratified splitting: an approach to train better classifiers*, arXiv preprint, 2020. https://doi.org/10.48550/arXiv.2010.06099
34. J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, *Boosting methods for multi-class imbalanced data classification: an experimental review*, J. Big Data **7** (2020). https://doi.org/10.1186/s40537-020-00349-y
35. M. Bader-El-Den, E. Teitei, and T. Perry, *Biased random forest for dealing with the class imbalance problem*, IEEE Trans. Neural Netw. Learn. Syst. **30** (2019), 2163–2172.

AUTHOR BIOGRAPHIES

P. Antony Seba received her BE degree in Computer Science and Engineering from Manonmaniam Sundaranar University, Tamil Nadu, India, and her ME degree in Computer Science and Engineering from Anna University, Tamil Nadu, India. She is currently a PhD Research Scholar at the Indian Institute of Information Technology Kottayam, Kerala, India. Her research interests are machine learning, big data analytics, and data science.

J. V. Bibal Benifa is presently associated with the Indian Institute of Information Technology Kottayam, Kerala, India, as an assistant professor. She obtained her BE, ME, and PhD degrees in the domain of cloud computing at Anna University, India. Her research interests are cloud computing, big data analytics, and machine learning. She is a life member of ISTE, India, and she is involved in developing social applications.

How to cite this article: P. A. Seba and J. V. B. Benifa, *Relevancy contemplation in medical data analytics and ranking of feature selection algorithms*, ETRI Journal **45** (2023), 448–461. https://doi.org/10.4218/etrij.2022-0018