Intelligent feature selection with modified K-nearest neighbor for kidney transplantation prediction

Dalia M. Atallah¹ · Mohammed Badawy² · Ayman El-Sayed²

© Springer Nature Switzerland AG 2019

Abstract

The prediction of kidney transplantation outcome is an important challenge and does not need emphasis because of the lack of available organs. Graft survival prediction is significant to help physicians to take the right decision and enhance survival rate by changing medical procedure. Also, it helps in the best choice of the existing kidney donor and the immunosuppressive management suitable for a patient. But the exact prediction of the graft survival is still not accurate despite of the advancements in this field. The purpose of our research is to design an intelligent kidney transplantation prediction method to solve the prediction problem by utilizing data mining methods. The novelty of this study is focused in presenting: (a) an integrated prediction method, (b) a new intelligent feature selection method, and (c) a modified K-nearest neighbor. Choosing the proper variables is accomplished by merging three feature selectors. The new proposed feature selection method is accomplished using gain ratio, naïve Bayes, and genetic algorithm. Next, the cleaned dataset is utilized to provide quick and precise outcome throughout a modified K-nearest neighbor classifier. Each stage of this proposed method has been evaluated using intense experiments. Experimental results demonstrate the efficiency of all the steps of the proposed method. Additionally, the proposed method has been evaluated versus latest methods. The results presented that this method outperformed all latest and similar literature methods. This method can as well be employed to other related transplant datasets.

Keywords Kidney transplantation · Graft failure · Gain ratio · Feature selection · Naïve Bayes · Genetic algorithm · K-nearest neighbor

1 Introduction

Data mining methods are used in many disciplines and have created a progress to people's life. Diseases affect people differently, and end-stage organ failure is one of the worst situations. It affects people due to hypertension, diabetes, and other causing diseases. In this case, the best solution is organ transplantation [1].

The prediction of kidney transplantation result is significant in helping physicians to take the right decision, saving patient's life, and make a good utilization of the available resources. The available kidneys are low and patient wait long time to find the suitable kidney. Some of the

kidney transplantation operations fail due to graft failure. The unsuitable match between both recipient and donor can cause this problem [2]. Additionally, the economic influence of being back to dialysis or search for another transplant.

Kidney transplant patients who have suffered from renal failure have better results than patients staying on a waiting list [3]. But, patient survival rate is lowered by graft failure and return patient to waiting list [4]. Despite of that kidney transplantation is the preferred management of renal failure, the number of the available kidneys for transplant is less than the number of patients [5]. Thus, efforts directed at enhancing graft survival interval are

✉ Dalia M. Atallah, daliaat@hotmail.com | ¹Urology and Nephrology Center, Mansoura University, Mansoura, Egypt. ²Computer Science and Engineering Department, Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt.

very significant for clinical practice. One way to enhance long-term allograft survival is to discover factors that negatively affect the result and observing the relations between and among them and planning a prediction method to enhance the result. Using models to predict allograft result may consequently make an important influence on physician plan management and improve complete clinical result [6].

Graft survival is the period of time that the patient does not require dialysis or one more transplant and the transplanted kidney still works [7]. All together if considerable papers have inspected the variables of kidney transplantation [7, 8], limited papers have employed data mining methods in designing a prediction method for kidney transplantation [9–11]. Latest papers have demonstrated that data mining methods offer improved outcomes in predicting graft survival [2].

Commonly, feature selection and outlier rejection are two essential matters that should be focused on precisely before concerning a prediction method as they have powerful influence on the method operation. Graft survival is influenced by a diversity of features and many useless features may perhaps exist in kidney transplantation data. Numerous prediction methods do not offer good results with big volumes of variables. Feature selection methods provide a great effect in raising the method accuracy along with offering quicker results by means of reducing the studied features to the just valuable ones. It also decreases the method complexity [12].

On the contrary, training the prediction method without data preparation, which often contains outliers, reduces the prediction method accuracy since the prediction has influenced by those unimportant instances whose achievement is very poor. Outlier rejection is an exceptional data mining procedure [13]. It offers an improvement to prediction accuracy because it helps to select the correct instances. Therefore, outlier rejection is important for rejecting all bad instances that affect the result.

This paper presents a novel kidney transplantation prediction method through data analysis. The novelty of this study is focused in presenting an integrated prediction method, a new intelligent feature selection, and a modified K-nearest neighbor. The proposed prediction method combines variable selection with outlier rejection and data mining techniques to provide enhanced predictive powers. This proposed methodology is also suitable for any transplant dataset.

This proposed method is introduced in three phases. These phases are (1) data organization phase (DOP), (2) variable selection method (VSM), and (3) outlier rejection and prediction phase (PP). During DOP, input data are pre-processed to be prepared for use in the next phases. VSM is used to select the reduced variables list to diminish the

method complexity and reduce the variables dimension to enhance the prediction method. Variable selection is accomplished utilizing gain ratio, naïve Bayes, and genetic algorithm. PP introduces a modification on K-nearest neighbor algorithm to classify testing instances, reject outliers, and predict the outcome of graft survival.

The validity of this proposed method is evaluated using urology and nephrology center dataset. Each stage of the proposed prediction method has been assessed throughout intense experiments. The evaluation results emphasize that our proposed method is efficient. The proposed VSM can select the reduced variables list efficiently. The proposed PP provides good results regarding prediction accuracy. Also, the proposed prediction method has been assessed versus similar literature, and the results outperform the recent and similar literature techniques in prediction metrics.

Our paper is designed as follows: Part 2 presents a background about variable selection, outlier rejection, and K-nearest neighbor algorithm. Part 3 presents an outline about the recent techniques used for predicting graft survival especially kidney transplantation. Part 4 presents the proposed method in detail. Part 5 presents the used dataset, the evaluation results, and discussion. Part 6 introduces the conclusion. Finally, part 7 introduces the used references.

2 Background

This part introduces the basic concepts used in this research. These basic concepts are variable selection, outlier rejection, and K-nearest neighbor algorithm.

2.1 Variable selection

Variable selection is an important method in constructing a prediction method [14]. The dataset may contain some unrelated variables. Hence, selecting the proper variable selection method that can specify the unrelated variables can have a positive effect in enhancing the result of machine learning techniques, preventing overfitting, producing better accuracy, and leading to quicker economical models [15].

There are three methods for feature selection which are filter, wrapper, and embedded methods [16–19]. Filter method assesses sets of variables by inspecting information content or measuring correlation between variables [12]. This method does not rely on the utilized data mining technique. Variable weights are computed, and the low weight variables are discarded [15]. Wrapper method employs data mining methods to evaluate the subsets of features using the resulting accuracy [12]. It utilizes the

outcomes of machine learning methods to specify the efficiency of the given subset of features [15]. Feature selection using embedded methods is accomplished during training of the best variables (i.e., neural network weights among the input and the hidden layer). Table 1 offers a comparison between variable selection methods.

2.2 Outlier rejection

Outlier recognition has been manipulated for periods to identify and eliminate irregular examinations from information. Outliers appear owing to mechanical errors, alternations in system behavior, person mistake, and tool fault or easily throughout normal abnormalities instances [20].

Outlier rejection is capable of discovering the little instances embedded in the dataset that affects the performance results [17, 21]. Therefore, these instances must be excluded from the dataset because these outliers reduce the predictive accuracy [13].

2.3 K-nearest neighbor algorithm

It is an easy algorithm and one of the oldest rules used for model classification [19]. K-nearest neighbor algorithm presents the training instances of the dataset as points in the variable space. These points split into numerous split objectives. To classify an instance P_i : (1) demonstrate the instance in the feature space. (2) Compute the spaces between the instance P_i and the K th nearest instances. (3) The instance P_i is classified along with the class with the highest selection of the neighbors. Figure 1 illustrates the algorithm supposing three objective classes (i), (ii), and (iii), and $K = 7$. Then, the point P_i will be classified to class (ii) as class (ii) contains the highest selection of the neighbors.

Fig. 1 K-nearest neighbor classifier example

3 Previous work

The importance of predicting kidney transplantation result does not require emphasis. Preoperative graft survival prediction might simplify physician evaluation and boost survival rate by altering medical method. Also, it allows the selection of the best presented donor with the best immunosuppressive management for a patient [10, 22].

Numerous approaches have been utilized to build predictive model. A multivariate analysis was used for kidney transplantation prediction from deceased donor [23]. In alternative study, multivariate analysis was employed for creatinine stages prediction in kidney recipients from living donors [24]. The possibility of graft survival for deceased donor was learned utilizing tree regression method [6, 25]. Artificial neural networks

Table 1 Comparison of variable selection methods [18]

Method	Advantage	Disadvantage
Filter method	Computationally simple	Every feature is calculated singly and therefore does not put into consideration the dependencies between the variables
	Very fast	Absence of the effect of variable dependencies on reducing the accuracy contrasted to other methods
	Independent on any algorithm	
	Has better complexity than any other method ($O(n)$)	
Wrapper method	Put in consideration the relation among the subsets of features and the selected method	Overfitting compared to other methods
	Feature dependencies are considered	The calculations are intensive
	Easy	
Embedded method	Cooperates with the classification method	Overfitting
	Better complexity compared to other methods	

were utilized to predict the probability of delayed graft function after kidney transplantation from deceased donors [26].

Kaplan–Meier survival was utilized by Rana et al. [27] to offer an impression of kidney transplantation survival. Numerous other papers have examined the influence of patient features (age, sex, and nation) [28] and preoperative transplant factor on graft status after transplant by utilizing the Kaplan–Meier survival function [29, 30]. These preceding educations were constrained by using few variables. Therefore, it is required to utilize greater dataset that contain a larger number of candidate predictors to discover possible unknown variables among the diverse predictors, which could influence the kidney transplant results.

Data mining methods have been employed widely in the medical area of organ transplantation. Atallah et al. [31] designed a new integrated prediction method to predict graft survival outcome. They used information gain merged with naïve Bayes to choose the important variables. Then, they used K-nearest neighbor to predict the graft survival. The results outperformed latest methods. Kusiak et al. [32] evaluated the efficiency of rough sets and two decision trees in forecasting the kidney dialysis patient's survival. While the outcomes of both approaches were precise, it is not well defined if this method operates to a bigger dataset. In extra study, decision trees were also utilized to forecast the acute liver failure prediction [30]. To widen the candidate predictors, another research paper utilized demographic, pharmaceutical, and clinical data to evaluate the accuracy of artificial neural network and nomogram to predict graft outcome at five years. The results revealed that artificial neural network outperformed nomogram [10]. In extra study, Lin et al. [22] concluded the influence of candidate predictors on kidney survival rate using logistic regression, Cox proportional hazard method, and artificial neural networks. Dag et al. [33] utilized four classification techniques to categorize graft outcome for heart patients. They utilized data balancing techniques to deal with the unbalanced class of the dataset. Their analysis revealed that logistic regression together with SMOTE provides the best result prediction. Delen et al. [34] utilized a database that has a lot of features using four machine learning methods, and they accomplished sensitivity analysis on the preferred method to obtain the significant predictors. In their analysis, support vector machines were the best predictor. Another study presents a weighted decision tree algorithm for graft survival prediction after kidney transplantation using patient's data before transplant and it can predict graft survival with good results [35]. Hence, it appears that the mutual idea between these studies is their capability to effectively predict a result utilizing a group of candidate predictors.

To create a precise prediction model, numerous requirements must be satisfied: The use of a dataset that introduces a large patient population to find unknown patterns between the variables that can change the clinical results [36]. Additionally, use of an intelligent variable selection technique that can specify the most significant variables in the dataset accurately and to promote the operation of the learning classifier. Lastly, the use of outlier's rejection that can identify the instances with exceptional behavior and reject it.

The purpose of our study is to design a prediction method that can classify graft status successfully. The novelty of this study is focused in presenting an integrated prediction method, a new intelligent feature selection, and a modified K-nearest neighbor. The proposed prediction method combines variable selection with outlier rejection and machine learning methods to enable better predictive abilities. Therefore, it is essential to employ a variable selection technique that can specify influential variables and in the same time eliminates the unimportant ones. Feature selection is accomplished by gain ratio, naïve Bayes, and genetic algorithm. This study introduces a method composed of three phases, namely: (i) data organization phase (DOP), (ii) variable selection method (VSM), and (iii) outlier rejection and prediction phase (PP).

4 The proposed method

The purpose of this paper is to design a novel proposed method to classify graft status outcome successfully. Figure 2 illustrates the proposed method. In the following sections, the three phases of this method will be explained in more details.

Fig. 2 Proposed method (PM)

4.1 Data organization phase (DOP)

In this phase, the dataset is pre-managed to be ready for building the model and testing it accordingly. This process is called data cleaning. It contains three phases. The first phase excludes all the variables that are related to the operative and postoperative procedures since the proposed method is planned to identify kidney transplantation result before the operation. The second phase excludes the variables that have no influence in the prediction procedure (e.g., the patient name, the id number of the hospital, and investigations dates). The third phase excludes from the dataset all the instances that contain missing data.

4.2 Variable selection method (VSM)

Commonly, variable selection is a significant procedure of any data mining procedure. Choosing the proper features will surely boost the method's prediction accuracy while reducing the time and processing penalties. This paper offers a new feature selection procedure for identifying the highly influential and informative variables to obtain the reduced variables list.

Genetic-based feature selection is proposed based on genetic algorithm. It utilizes a filter and a wrapper method with genetic algorithm to effectively choose the effective collection of features. Gain ratio is used as filter method and naïve Bayes as wrapper method. The filter method uses a function calculation for every feature to specify the best influential features in the dataset according to these calculations. The wrapper method makes evaluation of subsets of the features using cross-validation. The reduced variables list is selected based on this evaluation.

Gain ratio is used to measure the weight of each attribute. The idea of choosing gain ratio is because it overcomes the bias toward attributes with too many values. Information gain is a ranking variable selection technique [37]. It selects the features that have more information with respect to classes [38]. Information gain depends on entropy; it shows the information contribution of the attribute with the dataset.

Suppose that S is the group of n tuples and C is the group of K classes, $P(C_i, S)$ presents the part of the tuples in S that has class C_i . Then, the information introduced from this class is given by [39]:

$$\text{Info}(S) = - \sum_{i=1}^k P(C_i, S) * \log(P(C_i, S)) \quad (1)$$

Suppose S_i is the group of tuples and A_i is the attribute value A , then the knowledge extracted from class participation after partitioning by variable A ($\text{Info}_A(S)$) is introduced in Eq. (2) [39]:

$$\text{Info}_A(S) = - \sum_{i=1}^v \frac{|S_i|}{|S|} * \text{info}(S_i) \quad (2)$$

Hence, the difference between $\text{Info}(S)$ and $\text{Info}_A(S)$ introduces the information calculation by splitting S based on variable A as shown in Eq. (3) [39]:

$$\text{Gain}(A) = \text{Info}(S) - \text{Info}_A(S) \quad (3)$$

The information gain is influenced by attributes with many values. So gain ratio overcomes this problem by using split information as in Eq. (4) [40]:

$$\text{SplitInfo}_A(S) = - \sum_{i=1}^v \frac{|S_i|}{|S|} * \log_2 \left(\frac{|S_i|}{|S|} \right) \quad (4)$$

Then, the gain ratio is calculated as introduced in Eq. (5) [40]:

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}_A(S)} \quad (5)$$

Naïve Bayes is the most powerful probabilistic classifier as evidenced [41, 42] that uses Bayes' theorem to calculate class probability and classify the instances according to highest probability. The idea behind using naïve Bayes classifier is that it simplifies the computation and gives high accuracy and speed [40]. The way naïve Bayes operates can be specified as follows [40].

Assume D is group of instances with their known class labels. Every instance is represented by n -dimensional vector, $X = (X_1, X_2, \dots, X_n)$ with n variables ($V_1, V_2, \dots, V_n$). Assume that $(C_1, C_2, \dots, C_m)$ represents m classes. For any instance X , the classifier will predict the corresponding class by specifying the highest probability on X , i.e., the class C_j that has maximized $P(C_j|X)$. In other words, we maximize $P(C_j|X)$ only if [40]:

$$P(C_j|X) > P(C_i|X) \quad \text{for } 1 \leq j \leq m, \quad j \neq i$$

Then, Eq. (6) represents the Bayes theorem that calculates the $P(C_j|X)$ [40]:

$$P(C_j|X) = \frac{P(X|C_j)P(C_j)}{P(X)} \quad (6)$$

Genetic algorithm as a wrapper method is a famous random search and optimization method [43–45]. Its idea is from the human being evolution process [46]. It can be used in several optimization, explore, and data mining problems [47, 48]. Genetic algorithm works using iterative

manner to produce new generations from old ones using standard genetic operators for example selection, crossover, and mutation. It uses a fitness function to evaluate each generation [39]. Additional details on genetic algorithm are introduced in [48].

Figure 3 clarifies the proposed variable selection algorithm in detail. First, populations of individuals were initialized randomly. The generation is started by zero. The gain ratio is calculated for every feature in this population. The features that have gain ratio greater than zero are added to the reduced variables list. This list will be trained and tested manipulating naïve Bayes classifier. Basically, naïve Bayes examines the influence of every distinct variable and evaluates its significance on accuracy consequently. It eliminates one input variable from the input variables at a time, and the resulting variables are subsequently employed for training and testing by means of naïve Bayes classifier. The resulting naïve Bayes classifier accuracy is compared with the classifier accuracy built using all variables. If the classifier accuracy is decreased, then this variable will be added to the reduced variables list. Otherwise, it is discarded. Then, genetic algorithm produces another population iteratively using its policies, which are selection, crossover, and mutation [49, 50]. Selection specifies the parents that will be tested in the next generation. The next generation is selected randomly, and each variable is coded to binary zero or one. The generation length introduces the number of variables. The binary one shows a selected variable while the binary zero shows a non-selected variable. Typical crossover and mutation are applied to produce new offspring [39, 51]. This generation is evaluated again manipulating hybrid filter and wrapper process. The algorithm calculates the accuracy of each generation until the termination criteria are satisfied. In our proposed genetic-based feature selection algorithm, the algorithm stops when the number of iterations reached

Fig. 3 Proposed variable selection method

the maximum limit or when the reduced variables list (RVL) that have the highest accuracy between the generations is reached, and the accuracy did no change for several iterations. Algorithm 1 presents the proposed variable selection in detail.

Algorithm 1: The proposed variable selection method**Input:**

Input dataset

 $f_1, f_2, \dots, f_n$ is the features of the dataset where n is variables numbers.

Initialize high accuracy

Initialize Generation = 0

Output:

Reduced Variables List (RVL)

```

1. Begin
2. While (Generation < max. no. of iterations)
3.   Initialize randomly population of individuals
4.   For each feature ( $F_i$ ),  $i=1$  to no. of features do
5.     For  $k=1$  to  $m$ , where  $m$  is no. of instances do
6.       Calculate  $GR_k(F_i)$ 
7.     End for
8.     If  $GR_k(f_i) > 0$ 
9.       Add  $f_i$  to  $RVL_1$ 
10.    End if
11.  End for
12.  Train and test NB classifier using features in  $RVL_1$ 
13.  Calculate accuracy  $Acc$ 
14.  for each feature  $f_i$  in  $RVL_1$  do
15.    Remove  $f_i$  from  $RVL_1$ 
16.    train and test NB classifier using remaining features in  $RVL_1$ 
17.    Calculate accuracy  $Acc(f_i)$ 
18.    If  $Acc(f_i) < Acc$  then
19.      add  $f_i$  to  $RVL_2$ 
20.    End if
21.  End for
22.  If  $Acc(RVL_2) > high\ accuracy$ 
23.    Set high accuracy = accuracy of  $RVL$ 
24.  End if
25.  Update  $RVL$  if there is a better solution
26.  Create a new population by genetic operation
27.  Selection
28.  Crossover
29.  Mutation
30.  Generation = Generation + 1
31. End while
32. return  $RVL$ 
33. End

```

symbol	Meaning
GR	Gain ratio
NB	Naïve Bayes classifier
RVL	Reduced variables list
RVL_1	Reduced variables list from gain ratio filter method
RVL_2	Reduced variables list from naïve Bayes wrapper method

4.3 Outlier rejection and prediction phase (PM) using modified K-nearest neighbor algorithm

The usually utilized distance-based algorithm is K-nearest neighbor (KNN), which depends on distance measures as the only classification criteria. KNN is conceptually much close to the concept of predicting kidney transplantation outcomes through data analysis. Another benefit is that, being a lazy learner, KNN does not need learning and preserving a given model. So, the model can adapt to rapid modifications [52].

Modified K-nearest neighbor (MKNN) is a modification on the classical KNN classifier. Furthermore, to the distance measures, MKNN also puts in consideration the quality of each instance from the neighbors of the tested instances. As the test item may be near in distance to the instances at the edges of a class but it is much more similar in characteristics with instances in another class so the distance measure is not a satisfied criterion.

Algorithm 2 introduces the operation of MKNN algorithm. Initially, the training instances are demonstrated

in n -dimensional feature space. Then, the center of each class group is identified using Eq. (7):

$$C(i) = \left(\frac{1}{n} \sum_{i=0}^n x_i^1, \frac{1}{n} \sum_{i=0}^n x_i^2, \dots, \frac{1}{n} \sum_{i=0}^n x_i^m \right) \quad (7)$$

where C denotes the center of the class, x_i^m in the m th dimension of the i th instance, and n is the number of instances in the class. Next, the distance between the class center C and each class group, denoted as x_m , are computed for each class using Eq. (8):

$$x_m = \text{Dis}(y, z) = \sqrt{\sum_{i=1}^n (y_i - z_i)^2} \quad (8)$$

where $\text{Dis}(y, z)$ is the distance between instance y and z , n is the instances number, y_i is the i th dimension of instance

y , z_i is the i th dimension of instance z , and m is the class group number. Subsequently, Calculate the average of distances denoted as D_{th} . D_{th} is calculated as shown in Eq. (9):

$$D_{th} = \sum \frac{x_m}{m} \quad (9)$$

where m is the class group number.

Then, for each class group E_i , if the distance between the center of the class and the class group is less than or equal D_{th} then the class item E_i is one of the class member group. Otherwise, delete this point as outlier rejection.

Then, from the H points, pick the nearest M instances near the test point P_t . Finally, classification of each test point P_t is the class that have the highest similar of the M elected instances that P_t is belonging.

Algorithm 2: The modified K-nearest neighbor Algorithm

Input:

The used dataset

Output:

Classify test point

1. Begin

2. For every group of classes (G_L) do

3. The training instances are demonstrated in n dimensional feature space.

4. calculate the class group center $C(i)$ using equation (7)

$$C(i) = \left(\frac{1}{n} \sum_{i=0}^n x_i^1, \frac{1}{n} \sum_{i=0}^n x_i^2, \dots, \frac{1}{n} \sum_{i=0}^n x_i^m \right) \quad (7)$$

Where C presents the center of the class group, x_i^m is in the m^{th} dimension of the i^{th} instances, and n is the instances number in the group.

5. calculate the distance D among the group center $C(i)$ and each group item using equation (8)

$$x_m = \text{Dis}(y, z) = \sqrt{\sum_{i=1}^n (y_i - z_i)^2} \quad (8)$$

where $\text{Dis}(y, z)$ is the distance between instance y and z , n is the total instance number, y_i is the i^{th} dimension of instance y , z_i is the i^{th} dimension of instance z , m is the class group number.

6. Calculate the average of distances x_m denoted as D_{th} . D_{th} is calculated as shown in equation (9).

$$D_{th} = \sum \frac{x_m}{m} \quad (9)$$

where m is the class group number.

7. For each class item E_i do

8. If $D \leq D_{th}$ then

9. The class item E_i is one of the class member group.

10. Else

11. Delete class item E_i

12. End If

13. End For

14. From the H points, choose the nearest M items near the test point P_t .

15. Lastly, the test point P_t is classified regarding the class which has the highest similar of the M selected instances.

16. End For

17. End

An example is shown in Fig. 4. It introduces classes A, B, and C in a two-dimensional space. It is required to classify one event P_t using 7-nearest neighbor algorithm. The steps of the modified K-nearest neighbor algorithm are clarified in detail in Fig. 4.

5 Results and discussion

In this part, the principal contributions that were presented in the proposed method will be evaluated. Those contributions are (i) the new proposed variable selection method (VSM), (ii) the outlier rejection and prediction phase (PP) based on MKNN classifier, and (iii) the new proposed prediction method.

5.1 The used dataset

For implementing the proposed prediction method, the kidney transplantation dataset from urology and nephrology center, Mansoura University is used. The dataset included information on the kidney transplant such as demographic data, medical history, and medical conditions during and after the transplant. The dataset also contains some preoperative factors for both recipients and donors, in addition to other factors such as date of transplant, dialysis information. The required approvals of this study were obtained from the institutional review board (IRB) committee of Mansoura University (R/19.01.16).

Four decades ago, renal transplantation was started at urology department, Mansoura University, Egypt. The first renal transplantation was a patient that was suffering from

1- The training data is presented in the feature space.

2- Calculate the class center of each class and the distance between the center and each point in the class.

3- Calculate the average distance in each class.

4- Classify the target point.

Fig. 4 Steps of the proposed modified K-nearest neighbor

Table 2 Confusion matrix

	Actual class	
	Survival	Not survival
Predicted class		
Survival	True positive (TP)	False negative (FN)
Not survival	False positive (FP)	True negative (TN)

chronic pyelonephritis and was ended in renal failure. It was a successful operation. Therefore, renal transplantation operation accomplished was nearly 80 patients yearly. Urology and nephrology center gives health care for seven million populations. In this locate, seventeen hemodialysis centers produce service for nearly 2000 patients. So, there will be a long waiting list. The problem is that cadaveric transplantation is not legal in Egypt [53].

In March 2017, 2811 renal transplant were accomplished in the center. After deleting the instances that have missing values, 2750 patients were included in the experiments. The age of the patients was (mean $\pm$ SD, 29.1 $\pm$ 10.9), male to female proportion was 2026 (74%): 702 (26%), the age of the donors was (mean $\pm$ SD, 36.7 $\pm$ 10.2), and male to female proportion was 1246 (46%): 1482 (54%). The proportion between living related donors to living unrelated donors was 2293 (84%): 435 (16%). The proportion between the donors with same blood group to donors with different blood group was 2192 (80%): 536 (20%). Follow up for the graft was ranged from 0 to 33.5 years (mean $\pm$ SD, 7.7 $\pm$ 1.4).

5.2 Performance metrics

During the following experiments, five metrics will be evaluated: (i) accuracy, (ii) precision, (iii) sensitivity, (iv) *F*-measure, and (v) error. Confusion matrix is utilized to compute values of these metrics. Confusion matrix is illustrated in Table 2. The performance metrics are described as follows

True positive (TP) The instances number that the method predicts approves with historic cases.

True negative (TN) The instances number that the method predicts approves with historic cases.

False positive (FP) The instances number that the method predicts while historic cases were graft failure.

False negative (FN) The instances number that the method predicts while historic cases were graft survival.

Accuracy presented in Eq. (10) is considered as the proportion of the instances classified correctly to the instances number.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

Precision presented in Eq. (11) is considered as the percentage of the rate of the instances classified correctly in all detected instances.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (11)$$

Sensitivity presented in Eq. (12) is considered as the percentage of correct instances among all instances.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (12)$$

F-measure presented in Eq. (13)

$$F\text{-measure} = 2 * \frac{\text{precision} * \text{sensitivity}}{\text{precision} + \text{sensitivity}} \quad (13)$$

Error presented in Eq. (14)

$$\text{Error} = 1 - \text{accuracy} \quad (14)$$

5.3 Constraint settings

In the proposed variable selection method, gain ratio is used as a filter method. Each variable has a value which specifies the importance of the variable. The variable with zero value is discarded. The features that have gain ratio greater than zero are selected. Higher gain ratio introduces higher chance of having pure classes in the objective class. Also we use genetic algorithm to obtain the reduced variables list. The population is set randomly. The population consists of many chromosomes. Each one is coded by a binary string [54]. Bit value (0) denotes a non-chosen chromosome, and bit value (1) represents a selected chromosome. The size of the chromosome introduces the variables number in the population. Table 3 introduces the genetic algorithm factors used in the experiments.

In the proposed prediction method, MKNN is proposed for classification. We specified the best value of *k* experimentally. We start with *k* = 1 and test the error rate of the classifier. We

Table 3 Genetic algorithm factors

Factor	Value or method
Population size	100
Selection method	Roulette
Crossover rate	0.8
Mutation rate	0.2
Elite count	2

repeat this process many times by incrementing the k value to add one additional neighbor. We choose the k value that produces the highest accuracy with minimum error.

5.4 K-fold cross-validation

K -fold cross-validation is a method used to compare the performance metrics of the prediction method. It is used to minimize the bias between the random training and testing samples [55, 56]. The dataset is separated to k mutually equal size groups instead of dividing the dataset into two random sampling groups (training group and testing group) that are prone to be biased. In this method, the prediction method is tested k times by using the k test sets. The performance metric is the k performance metrics average of the k folds. Equation 15 presents the performance measure [57]:

$$PM = \frac{1}{K} \sum_{i=1}^k PM_i \quad (15)$$

where PM specifies the performance metric and k is the fold's number. In our experiments, we use ten-fold cross validation. Figure 5 introduce an explanatory figure for ten-fold cross validation. The unfilled sections in the figure introduce the training sets while the filled sections introduce the testing sets.

5.5 Evaluating the proposed variable selection method

This study introduces a new novel intelligent variable selection method that combines between filter and wrapper methods. It has used gain ratio, naïve Bayes, and genetic algorithm. Naïve Bayes has been verified as the highly professional probabilistic classifier [41, 42], so we will use it to verify our proposal efficiency. First, all

variables in the dataset are used to train naïve Bayes classifier and are called without feature selection (WFS). Second, we trained the steps of our proposed variable selection method using NB Classifier. The steps of our method are called gain ratio (GR), naïve Bayes (NB), genetic algorithm (GA). Finally, we trained the proposed variable selection method using NB classifier. Experimental results are introduced in Figs. 6, 7, 8, 9 and 10. In general, the performance

Fig. 6 Accuracy of the different feature selection methods

Fig. 7 Precision of the different feature selection methods

Fold 1									
Fold 2									
Fold3									
Fold 4									
Fold 5									
Fold 6									
Fold 7									
Fold 8									
Fold 9									
Fold 10									

Fig. 5 Explanatory figure for ten-fold cross validation

Fig. 8 Sensitivity of the different feature selection methods

Fig. 9 F-measure of the different feature selection methods

Fig. 10 Error of the different feature selection methods

metrics increased by increasing the instances number. The highest number of instances presents the highest performance metrics in maximizing the accuracy, precision, sensitivity, and F-measure while minimizing the error. The proposed variable selection method introduces the best performance metrics compared to other methods with the scores of 0.77, 0.78, 0.77, 0.77, and 0.22 for accuracy, precision, sensitivity, F-measure, and error respectively. Our proposed variable selection method merges the benefit of both gain ratio, naïve Bayes, and genetic algorithm

Fig. 11 Accuracy of the different classification techniques without feature selection

Fig. 12 Precision of the different classification techniques without feature selection

Fig. 13 Sensitivity of the different classification techniques without feature selection

as every one of them reward the weakness of the other. The results verified that the proposed variable selection method can choose the essential and influential variables to produce the reduced variables list. This stimulates the performance of the proposed variable selection method and diminishes its error.

Fig. 14 *F*-measure of the different classification techniques without feature selection

Fig. 15 Error of the different classification techniques without feature selection

5.6 Evaluating the proposed outlier detection and prediction phase

In this subsection, we will evaluate the proposed classifier (MKNN) with the classical KNN classifier and with other classifiers, namely decision tree (J48) and naïve Bayes (NB) without applying any feature selection. Figures 11, 12, 13, 14 and 15 show the accuracy, precision, sensitivity, *F*-measure, and error of applying the different classification methods. As shown in Figs. 11, 12, 13, 14 and 15, the highest accuracy, precision, sensitivity, *F*-measure are detected by the proposed MKNN because of rejecting the effect of outlier's instances that reduce the performance. It also achieves the least error competed to other classification techniques.

MKNN is a modification on KNN classifier. It uses distance based method to reject outliers and enhances the performance. Therefore, it promotes the advantages of KNN classifier. The proposed outlier rejection and prediction phase outperform all other techniques with the scores of 0.79, 0.8, 0.79, 0.79, and 0.25 for accuracy, precision, sensitivity, *F*-measure, and error, respectively.

Table 4 Most recent prediction methods utilized for evaluation

Author	Method	
	Feature selection	Classifier
Dag et al. [33]	–	Logistic regression
Topuz et al. [2]	Data analytic methods	Bayesian belief network
Oztekin et al. [1]	Genetic algorithm	Support vector machine
Atallah et al. [31]	IGBFS/NBBFS	K-nearest neighbor
The proposed prediction method	VSM	Modified K-nearest neighbor

Fig. 16 Accuracy of the proposed method against recent techniques

Fig. 17 Precision of the proposed method against recent techniques

5.7 Evaluating the proposed method

Finally, the proposed prediction method with all stages is examined against the most recent techniques used to design a prediction method to prove the efficiency of our proposed method. Table 4 introduces the most recent techniques used for evaluation and the method

Fig. 18 Sensitivity of the proposed method against recent techniques

Fig. 21 ROC area of the proposed method against recent techniques

Fig. 19 F-measure of the proposed method against recent techniques

Fig. 20 Error of the proposed method against recent techniques

used. Results are present in Figs. 16, 17, 18, 19, 20 and 21 and Tables 5, 6, 7, 8, 9, and 10. As demonstrated in Figs. 16, 17, 18, 19 and 20 and Tables 5, 6, 7, 8, 9 and 10, the proposed method presents the highest performance as it presents the highest accuracy among all recent

Table 5 Accuracy of the proposed method against recent techniques

Technique	No. of training dataset				
	550	1100	1650	2200	2750
Dag et al. [33]	0.52	0.54	0.57	0.66	0.71
Topuz et al. [2]	0.51	0.59	0.61	0.67	0.72
Oztekin et al. [1]	0.54	0.56	0.59	0.67	0.72
Atallah et al. [31]	0.62	0.69	0.73	0.78	0.80
The proposed prediction method	0.71	0.77	0.79	0.83	0.86

Table 6 Precision of the proposed method against recent techniques

Technique	No. of training dataset				
	550	1100	1650	2200	2750
Dag et al. [33]	0.52	0.54	0.57	0.65	0.71
Topuz et al. [2]	0.49	0.58	0.63	0.7	0.7
Oztekin et al. [1]	0.53	0.56	0.59	0.66	0.72
Atallah et al. [31]	0.52	0.61	0.61	0.65	0.69
The proposed prediction method	0.63	0.7	0.66	0.72	0.77

Table 7 Sensitivity of the proposed method against recent techniques

Technique	No. of training dataset				
	550	1100	1650	2200	2750
Dag et al. [33]	0.52	0.54	0.57	0.66	0.71
Topuz et al. [2]	0.52	0.59	0.61	0.67	0.72
Oztekin et al. [1]	0.54	0.56	0.59	0.67	0.72
Atallah et al. [31]	0.62	0.7	0.73	0.78	0.8
The proposed prediction method	0.71	0.77	0.79	0.83	0.86

Table 8 *F*-measure of the proposed method against recent techniques

Technique	No. of training dataset				
	550	1100	1650	2200	2750
Dag et al. [33]	0.52	0.54	0.57	0.66	0.71
Topuz et al. [2]	0.5	0.58	0.61	0.68	0.73
Oztekin et al. [1]	0.54	0.56	0.59	0.66	0.71
Atallah et al. [31]	0.54	0.62	0.65	0.69	0.73
The proposed prediction method	0.62	0.68	0.71	0.76	0.79

Table 9 Error of the proposed method against recent techniques

Technique	No. of training dataset				
	550	1100	1650	2200	2750
Dag et al. [33]	0.48	0.47	0.46	0.38	0.32
Topuz et al. [2]	0.49	0.45	0.39	0.33	0.28
Oztekin et al. [1]	0.46	0.44	0.41	0.34	0.28
Atallah et al. [31]	0.46	0.4	0.36	0.33	0.29
The proposed prediction method	0.41	0.34	0.31	0.27	0.23

Table 10 ROC area of the proposed method against recent techniques

Technique	No. of training dataset				
	550	1100	1650	2200	2750
Dag et al. [33]	0.52	0.54	0.59	0.68	0.74
Topuz et al. [2]	0.48	0.58	0.65	0.73	0.79
Oztekin et al. [1]	0.52	0.55	0.59	0.63	0.66
Atallah et al. [31]	0.6	0.68	0.71	0.76	0.78
The proposed prediction method	0.71	0.75	0.78	0.8	0.82

techniques with the score of 0.86. It also has the highest precision, sensitivity, *F*-measure, and ROC area with the scores of 0.77, 0.86, 0.79 and 0.82, respectively. Also, it has the minimum error with the score of 0.23. This proves the efficiency of the proposed method and all its stages.

6 Conclusion

The prediction of kidney transplantation result is very important and does not require emphasis. Hence, successful prediction method is an essential task. In this study, we have designed a new proposed method to classify graft result. This study introduced a method

composed of three stages, namely: (i) data organization phase (DOP), (ii) variable selection method (VSM), and (iii) outlier rejection and prediction phase (PP). The proposed method combines variable selection with outlier rejection and machine learning methods to enable better predictive abilities. The proposed prediction method includes new intelligent feature selection procedure and a modified K-nearest neighbor. The new proposed feature selection procedure collects between gain ratio, naïve Bayes, and genetic algorithm, which chooses the essential features from the dataset. Additionally, the proposed modified K-nearest neighbor introduced the outlier rejection and prediction module that used distance based measures with K-nearest neighbor to classify patients. The efficiency of the proposed method is evaluated using urology and nephrology center dataset. Each stage of the proposed method has been assessed throughout intense experiments. The overall method is examined to verify the compatibility of the proposed method. The evaluation results emphasize that our proposed method is efficient. The proposed VSM can select the reduced variables list efficiently. Results likewise specify the efficiency of the proposed outlier rejection and prediction phase as it enhances the prediction accuracy. Experimental results presented that the proposed method gave more precise results than most recent methods. In general, the results offered a new method that could benefit in progress the result of kidney transplantation. This method can as well be employed to other related transplant datasets.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Oztekin A, Al-Ebbini L, Sevkli Z, Delen D (2018) A decision analytic approach to predicting quality of life for lung transplant recipients: a hybrid genetic algorithms-based methodology. *Eur J Oper Res* 266(2):639–651
- Topuz K, Zengul FD, Dag A, Almekhi A, Yildirim MB (2018) Predicting graft survival among kidney transplant recipients: a Bayesian decision support model. *Decis Support Syst* 106:97–109
- Ojo AO, Hanson JA, Meier-Kriesche H-U, Okechukwu CN, Wolfe RA, Leichtman AB, Agodoa LY, Kaplan B, Port FK (2001) Survival in recipients of marginal cadaveric donor kidneys compared with other recipients and wait-listed transplant candidates. *J Am Soc Nephrol* 12(3):589–597
- Ojo AO, Wolfe RA, Agodoa LY, Held PJ, Port FK, Leavey SF, Calzard SE, Dickinson DM, Schmodder RL, Leichtman AB (1998)

- Prognosis after primary renal transplant failure and the beneficial effects of repeat transplantation: multivariate analyses from the United States renal data system 1, 2. *Transplantation* 66(12):1651–1659
5. Procurement O (2015) Organ procurement and transplantation network, vol 9. HRSA, DHHS, pp 36–42
 6. Krikov S, Khan A, Baird BC, Barenbaum LL, Leviatov A, Koford JK, Goldfarb-Rumyantzev AS (2007) Predicting kidney transplant survival using tree-based modeling. *ASAIO J* 53(5):592–600
 7. Hariharan S, Johnson CP, Bresnahan BA, Taranto SE, McIntosh MJ, Stablein D (2000) Improved graft survival after renal transplantation in the United States, 1988 to 1996. *N Engl J Med* 342(9):605–612
 8. Hoot N, Aronsky D (2005) Using Bayesian networks to predict survival of liver transplant patients. In: *Proceedings AMIA 2005 symposium*. American Medical Informatics Association, pp 345–349
 9. Brown TS, Elster EA, Stevens K, Graybill JC, Gillern S, Phinney S, Salifu MO, Jindal RM (2012) Bayesian modeling of pretransplant variables accurately predicts kidney graft survival. *Am J Nephrol* 36(6):561–569
 10. Akl A, Ismail AM, Ghoneim M (2008) Prediction of graft survival of living-donor kidney transplantation: nomograms or artificial neural networks? *Transplantation* 86(10):1401–1406
 11. Dag A, Topuz K, Oztekin A, Bulur S, Megahed FM (2016) A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival. *Decis Support Syst* 86:1–12
 12. Sajadfar N, Ma Y (2015) A hybrid cost estimation framework based on feature-oriented data mining approach. *Adv Eng Inform* 29(3):633–647
 13. Angiulli F, Pizzuti C (2002) Fast outlier detection in high dimensional spaces. In: *European conference on principles of data mining and knowledge discovery*. Springer, pp 15–27
 14. Aggarwal M (2013) Performance analysis of different feature selection methods in intrusion detection. *Int J Sci Technol Res* 2(6):225–231
 15. Blum AL, Rivest RL (1993) Training a 3-node neural network is NP-complete. In: *Machine learning: from theory to applications*. Springer, pp 9–28
 16. Zhang M, Yao J (2004) A rough sets based approach to feature selection. In: *IEEE annual meeting of the fuzzy information, 2004. Processing NAFIPS'04*. IEEE, pp 434–439
 17. Hung Y (2009) A neural network classifier with rough set-based feature selection to classify multiclass IC package products. *Adv Eng Inform* 23(3):348–357
 18. Khan M, Quadri S (2013) Effect of using filter based feature selection on performance of machine learners using different datasets. *BVICAM's Int J Inf Technol* 5:597–603
 19. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. *J Mach Learn Res* 10:207–244
 20. Hodge V, Austin J (2004) A survey of outlier detection methodologies. *Artif Intell Rev* 22(2):85–126
 21. Suh NP (2005) *Complexity: theory and applications*. Oxford University Press on Demand, Oxford
 22. Lin RS, Horn SD, Hurdle JF, Goldfarb-Rumyantzev AS (2008) Single and multiple time-point prediction models in kidney transplant outcomes. *J Biomed Inform* 41(6):944–952
 23. Poli F, Scalapogna M, Cardillo M, Porta E, Sirchia G (2000) An algorithm for cadaver kidney allocation based on a multivariate analysis of factors impacting on cadaver kidney graft survival and function. *Transpl Int* 13(1):S259–S262
 24. Zapletal C, Lorenz M, Woeste G, Wullstein C, Golling M, Bechstein W (2004) Predicting creatinine clearance by a simple formula following live-donor kidney transplantation. *Transpl Int* 17(9):490–494
 25. Goldfarb-Rumyantzev AS, Scandling JD, Pappas L, Smout RJ, Horn S (2003) Prediction of 3-yr cadaveric graft survival based on pre-transplant variables in a large national dataset. *Clin Transpl* 17(6):485–497
 26. Brier ME, Ray PC, Klein JB (2003) Prediction of delayed renal allograft function using an artificial neural network. *Nephrol Dial Transpl* 18(12):2655–2659
 27. Rana A, Gruessner A, Agopian VG, Khalpey Z, Riaz IB, Kaplan B, Halazun KJ, Busuttil RW, Gruessner RW (2015) Survival benefit of solid-organ transplant in the United States. *JAMA Surg* 150(3):252–259
 28. Heldal K, Hartmann A, Grootendorst DC, de Jager DJ, Leivestad T, Foss A, Midtvedt K (2009) Benefit of kidney transplantation beyond 70 years of age. *Nephrol Dial Transpl* 25(5):1680–1687
 29. Port FK, Bragg-Gresham JL, Metzger RA, Dykstra DM, Gillespie BW, Young EW, Delmonico FL, Wynn JJ, Merion RM, Wolfe RA (2002) Donor characteristics associated with reduced graft survival: an approach to expanding the pool of kidney donors 1. *Transplantation* 74(9):1281–1286
 30. Nakayama N, Oketani M, Kawamura Y, Inao M, Nagoshi S, Fujiwara K, Tsubouchi H, Mochida S (2012) Algorithm to determine the outcome of patients with acute liver failure: a data-mining analysis using decision trees. *J Gastroenterol* 47(6):664–677
 31. Atallah DM, Badawy M, El-Sayed A, Ghoneim MA (2019) Predicting kidney transplantation outcome based on hybrid feature selection and KNN classifier. *Multimed Tools Appl* 78(14):20383–20407
 32. Kusiak A, Dixon B, Shah S (2005) Predicting survival time for kidney dialysis patients: a data mining approach. *Comput Biol Med* 35(4):311–327
 33. Dag A, Oztekin A, Yucel A, Bulur S, Megahed FM (2017) Predicting heart transplantation outcomes through data analytics. *Decis Support Syst* 94:42–52
 34. Delen D, Oztekin A, Tomak L (2012) An analytic approach to better understanding and management of coronary surgeries. *Decis Support Syst* 52(3):698–705
 35. Atallah DM, Eldesoky AI, Amira Y, Ghoneim MA (2014) One-year renal graft survival prediction using a weighted decision tree classifier. *Int J Eng Technol* 3(3):327
 36. Kattan MW (2005) When and how to use informatics tools in caring for urologic patients. *Nat Rev Urol* 2(4):183
 37. Martín-Valdivia MT, Díaz-Galiano MC, Montejó-Raez A, Ureña-López L (2008) Using information gain to improve multimodal information retrieval systems. *Inf Process Manage* 44(3):1146–1158
 38. Mukras R, Wiratunga N, Lothian R, Chakraborti S, Harper D (2007) Information gain feature selection for ordinal text classification using probability re-distribution. In: *Proceedings of the Textlink workshop at IJCAI*. p 16
 39. Yang C-H, Chuang L-Y, Yang CH (2010) IG-GA: a hybrid filter/wrapper method for feature selection of microarray data. *J Med Biol Eng* 30(1):23–28
 40. Han J, Pei J, Kamber M (2011) *Data mining: concepts and techniques*. Elsevier, Amsterdam
 41. Inza I, Larrañaga P, Etxeberria R, Sierra B (2000) Feature subset selection by Bayesian network-based optimization. *Artif Intell* 123(1–2):157–184
 42. Qiang G (2010) An effective algorithm for improving the performance of Naïve Bayes for text classification. In: *2010 Second international conference on computer research and development*
 43. Holland JH (1975) *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press, Ann Arbor

44. Abed MA, Ismail AN, Hazi ZM (2010) Pattern recognition using genetic algorithm. *Int J Comput Electr Eng* 2(3):583
45. Tan F, Fu X, Zhang Y, Bourgeois AG (2008) A genetic algorithm-based method for feature subset selection. *Soft Comput* 12(2):111–120
46. Shahamat H, Pouyan AA (2015) Feature selection using genetic algorithm for classification of schizophrenia using fMRI data. *J AI and Data Min* 3(1):30–37
47. Golberg DE (1989) Genetic algorithms in search, optimization, and machine learning. Addison Wesley, Reading
48. Holland JH (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT Press, Cambridge
49. Ghareb AS, Bakar AA, Hamdan AR (2016) Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Syst Appl* 49:31–47
50. Pakath R, Zaveri JS (1995) Specifying critical inputs in a genetic algorithm-driven decision support system: an automated facility. *Decis Sci* 26(6):749–771
51. Ammu P, Preeja V (2013) Review on feature selection techniques of DNA microarray data. *Int J Comput Appl* 61(12):39–44
52. Hmeidi I, Hawashin B, El-Qawasmeh E (2008) Performance of KNN and SVM classifiers on full word Arabic articles. *Adv Eng Inform* 22(1):106–111
53. Ghoneim MA, Bakr MA, Refaie AF, Akl AI, Shokeir AA, Shehab El-Dein AB, Ammar HM, Ismail AM, Sheashaa HA (2013) Factors affecting graft survival among patients receiving kidneys from live donors: a single-center experience. *BioMed Res Int* 2013:1–9
54. Refaeilzadeh P, Tang L, Liu H (2009) Cross-validation. In: *Encyclopedia of database systems*. Springer, pp 532–538
55. Garson GD (1998) Neural networks: an introductory guide for social scientists. Sage, Thousand Oaks
56. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*, vol 2. Montreal, Canada, pp 1137–1145
57. Olson DL, Delen D (2008) Advanced data mining techniques. Springer, Berlin

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.